TAFER: AUDITING SELECTIVE REPAIR UNDER EVIDENCE SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Language-model outputs can match task targets without being supported by the available evidence. We introduce TAFER, a framework for auditing selective repair under evidence shift. TAFER separates answer-text correctness, answer-or-abstain decisions, and support assessments tied to explicit reference rules. In public fixed-pool comparisons, its deterministic selector chooses the lowest recorded-cost candidate that passes structural citation checks, or abstains when none passes. Shared-pool baselines distinguish feasibility filtering from ranking, while matched evidence interventions examine how removing cited or uncited evidence changes answering behavior. Finite-pool accounting separates candidate unavailability, certificate exclusion, and selection error; risk-coverage analysis makes abstention costs explicit. Results are heterogeneous: changes in a support proxy do not alone establish better task utility, and filtering gains need not imply ranking gains. Behavioral diagnostics distinguish increased abstention from continued answering after evidence removal. Natural-language intervention support uses operational verifier judgments, while candidate-pool support uses a gold citation-ID relation. Neither is treated as human support truth. These analyses characterize selective repair within tested protocols without conflating correctness, support, and abstention.

1 INTRODUCTION

Answer-level accuracy does not establish that a generated answer follows from the evidence shown to the model. A model may recover the right answer from memorization, a spurious cue, or an unsupported shortcut even after a supporting passage has been removed. Accuracy-only evaluation therefore turns some unsupported outputs green (Maynez et al., 2020; Rashkin et al., 2023).

We study a *false-green outcome*: a task-correct, non-abstaining answer unsupported by the visible evidence under a stated evaluation relation. We ask two questions. Does removing cited evidence change an answer or trigger abstention more often than equal-count uncited removal? Given a finite repair pool, does failure arise from missing good candidates, certificate exclusion, or selection after feasibility? Answer-string change is not automatically semantic error, and an operational certificate is not human truth.

TAFER audits selective repair from bounded candidate pools. In public fixed-pool comparisons, its deterministic policy selects the lowest recorded-cost candidate passing structural citation checks, and abstains when none passes; these checks do not test entailment. Controlled Stage3 uses a distinct family-conditioned selector and is a condition-aware diagnostic, not an equal-information deployment comparison (Appendix A). Raw generation, self-correction, abstention controls, random feasible selection, cost-only selection, and TAFER operate on shared pools, with evaluation metadata joined after selection. We report coverage and task-policy success alongside support proxies; task-policy success counts a task-correct answer admissible under the named evaluation relation or a contract-correct abstention, so blanket abstention cannot masquerade as repair.

Historical operational-judge decompositions are heterogeneous: FEVER's support-loss contrast is largely abstention-driven, while HotpotQA shows both abstention and continued-assertion components. In FEVER, same-key filtering improves the legacy composite while reducing coverage and correct-answer yield; comparison with analytic uniform-feasible selection does not establish an incremental ranking benefit (Appendix R.1). The main results report these endpoints separately; citation-ID-admissible yield is not correct-answer yield.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Our contributions are:

- a paired cited-versus-uncited removal analysis that separates response modes and objective task correctness;
- shared-pool comparisons that separate feasibility filtering, coverage costs, and within-feasible selection, retaining null and reversed contrasts; and
- a bounded regime map that preserves favorable, unresolved, and adverse outcomes across datasets, including the negative result for the tested Mistral configuration.

Evidence deletion, abstention, and candidate ranking are established tools. Our contribution is to measure response modes under matched evidence controls and to separate candidate feasibility from selection within shared pools. Reporting task correctness, checker output, coverage, and selector outcome separately prevents proxy or filtering gains from being mistaken for human-valid support or ranking benefit. Findings remain descriptive within tested protocols and declared evaluation relations.

2 Related Work

Evidence-grounded generation and revision. Retrieval-augmented generation supplies documents but does not ensure that generated claims follow from them (Lewis et al., 2020; Niu et al., 2024). ALCE evaluates citation-backed generation (Gao et al., 2023b); RARR finds evidence and post-edits unsupported text while preserving the original response (Gao et al., 2023a). TAFER shares this preservation objective, but audits selection within a fixed repair pool rather than retrieval-and-revision as a whole. FActScore, RAGAS, and ARES provide complementary factuality or RAG-evaluation pipelines (Min et al., 2023; Es et al., 2024; Saad-Falcon et al., 2024).

Context attribution and ablation. ContextCite attributes generations to context; SelfCite uses cited-context ablation for citation necessity and sufficiency (Cohen-Wang et al., 2024; Chuang et al., 2025). GroundLM likewise studies fixed-answer citation-set expansion, trading off support-ID coverage against distractors while cautioning that coverage is not claim-level support (Li, 2026). TAFER shares this concern but measures answer/abstain changes under evidence interventions and contrasts filtering with ranking within shared pools. Thus, ablation is diagnostic, not our novelty. Sufficient Context and C-RAG study context adequacy and certified generation risk; our probes establish neither semantic support nor distributional guarantees (Joren et al., 2025; Kang et al., 2024).

Attribution evaluation and refusal. AttributionBench benchmarks automatic attribution evaluation and documents its errors (Li et al., 2024); our two-judge sensitivity likewise tests evaluation dependence, not a new attribution benchmark. RefChecker extracts claim triplets and checks them against reference context (Hu et al., 2024); our operational checker is not a validated replacement. Attribute or Abstain evaluates attribution in long-document assistance (Buchmann et al., 2024). We also retain refusal, but study response modes and fixed-pool repair rather than long-document attribution performance. OR-Bench measures over-refusal (Cui et al., 2025). Checker agreement cannot itself establish human support truth.

Self-checking and correction. Sampling consistency, verification questions, and tool feedback can guide checking and revision (Manakul et al., 2023; Dhuliawala et al., 2024; Gou et al., 2024); intrinsic self-correction can fail without reliable feedback (Huang et al., 2024; Pan et al., 2024). We retain self-correction alongside coverage-matched abstention, verifier-only filtering, best-of-K, random feasible, cost-only, and RARR-like controls. A shared-pool RARR-like baseline is not a reproduction of the full RARR system.

Evidence-supervised stress tests. HotpotQA supplies sentence-level support for multi-hop questions; FEVER supplies support, refutation, and not-enough-information labels; SciFact supplies evidence rationales (Yang et al., 2018; Thorne et al., 2018; Wadden et al., 2020). Contrast sets and dynamic benchmarking motivate local, controlled comparisons (Gardner et al., 2020; Kiela et al., 2021). A transformation contract specifies an intended comparison; it does not establish support truth or make retrospective analysis confirmatory.

Selective and distributionally robust risk. Selective prediction evaluates risk with coverage (Geifman & El-Yaniv, 2017; Angelopoulos et al., 2024). Mechanistic abstention, context-adaptive abstention policies, and conformal tail-risk control pursue related error-reduction goals (Hedström et al., 2025; Tayebati et al., 2025; Chen et al., 2025). TAFER instead audits observed response modes and fixed-pool choices; it inherits no conformal guarantee. Group-robust objectives address failures hidden by averages (Sagawa et al., 2020; Liu et al., 2021; Zhai et al., 2021), and CVaR summarizes bounded-loss tails (Rockafellar & Uryasev, 2000). We use descriptive finite-protocol summaries and state the assumptions for zero-event binomial bounds (Clopper & Pearson, 1934; Brown et al., 2001). Finite tail concentration is not a population power law.

3 Problem Setup

An evaluation record is $x = (q, E, y, G, c)$: query q, visible evidence E, task oracle y, support relation G, and shift contract c. Only the prompt-safe query and evidence reach generation; y, G, c are evaluation metadata. For an output o, let

$$A(o, y) = \mathbb{1}[\mathrm{ans}(o) = y], \tag{1}$$

$$S(o, E, G) = \mathbb{1}[o \text{ is supported by } E \text{ under } G], \tag{2}$$

$$B(o) = \mathbb{1}[o \text{ abstains}]. \tag{3}$$

The false-green loss is

$$L_{\mathrm{FG}}(o; x) = A(o, y)(1 - S(o, E, G))(1 - B(o)). \tag{4}$$

For a policy π, unconditional false-green risk is $R_{\mathrm{FG}}(\pi) = \mathbb{E}[L_{\mathrm{FG}}(\pi(x); x)]$. Coverage and selective risk are

$$\mathrm{Coverage}(\pi) = \Pr[B(\pi(x)) = 0], \qquad R^{\mathrm{sel}}_{\mathrm{FG}}(\pi) = \begin{cases} \Pr[L_{\mathrm{FG}} = 1 \mid B = 0], & \mathrm{Coverage}(\pi) > 0, \\ 0, & \mathrm{Coverage}(\pi) = 0. \end{cases} \tag{5}$$

The zero-coverage convention makes the metric total-valued; an always-abstain policy still attains zero FGR with zero coverage and is therefore a degeneracy check rather than a successful repair method. Empirically, the definitions give the exact decomposition

$$R_{\mathrm{FG}}(\pi) = \mathrm{Coverage}(\pi) R^{\mathrm{sel}}_{\mathrm{FG}}(\pi). \tag{6}$$

The decomposition distinguishes safer answered predictions from risk reduction caused by lower answer coverage.

Support reference and operational checker. Let $H_i \in \{0, 1\}$ denote a determinate human support judgment when available; uncertain and not-applicable ratings remain outside this binary encoding. Let $C_i \in \{0, 1\}$ denote satisfaction of a complete, task-specific explicit support contract when one is available. Human judgments and contract-defined outcomes are distinct reference routes; neither is substituted for the other. Let $V_i \in \{0, 1\}$ denote the operational support judge. It is neither support ground truth nor the selection-feasibility predicate F defined below. Natural-language support in this study is not independently human-calibrated: substituting V for S yields checker-FGR, not human false-green risk. Legacy citation-subset and gold-ID-coverage scores are distinct ID rules (Appendix L); neither establishes semantic support. Programmatic truth applies only to the complete rule task.

For task-oriented tables, let $N(o; x)$ be the evaluation-contract indicator that abstention is correct and define

$$T(o; x) = (1 - B(o))A(o, y)S(o, E, G) + B(o)N(o; x). \tag{7}$$

We call $\mathbb{E}[T]$ task-policy success: it counts a task-correct answer supported under the declared relation G or a contract-correct abstention; it is not ordinary answer accuracy. It can therefore exceed answered coverage.

Evidence-shift contracts. We partition transformations into oracle-preserving shifts $\mathcal{T}_P$, oracle-changing shifts $\mathcal{T}_C$ with a constructed new oracle, and insufficient-evidence shifts $\mathcal{T}_I$ for which abstention is the correct action. The audited HF prompt uses question and visible evidence text; its

parser filters citations by visible IDs. Stage3 repair reads the explicit family, and its family-qualified row ID seeds candidate construction. Current-source replay matches configured $K = 4$ decision fields on 4,200/4,200 rows, but certificate relation labels differ on 2,450; the historical remote source hash is unavailable (Appendix A).

Response-mode estimand. For a complete full/cited/uncited triple, write B_i^t for abstention and V_i^t for the support proxy of a non-abstaining response. The loss from an initially proxy-supported answer is

$$L_i^t = \mathbb{1}[B_i^0 = 0, V_i^0 = 1]\big(\mathbb{1}[B_i^t = 1] + \mathbb{1}[B_i^t = 0, V_i^t = 0]\big), \tag{8}$$

where $t = c, u$ denotes cited or matched uncited removal. We report $L_i^c - L_i^u$ and its two components on the same complete-triple denominator. Initially unsupported or abstaining responses contribute zero; missing required checks remain unknown. Unlike L_{FG}, this diagnostic does not condition on answer correctness. It measures response modes, not false-green repair efficacy, causality, or human validity.

Finite-sample uncertainty. Exact one-sided Clopper–Pearson intervals require an independent-unit sampling model. Under that model, the zero-event upper limit is $1 - 0.05^{1/n}$. Expanded method–instance trials alone do not establish this independence. We report the observed benchmark frequency and do not infer a population upper limit from the expanded denominator.

4 Certificate-Constrained Selective Repair

Given an output o, TAFER generates a bounded pool $\mathcal{Z}_K(q, E, o)$ from prompt-safe inputs. Let $F(z, E) = 1$ mean that z is non-abstaining and passes both the structural verifier and certificate checks. This selection predicate is distinct from the behavioral support judge V. Each candidate has recorded cost $C(o, z) \geq 0$, assigned by type in the recovered public pools, not measured edit distance. The selector minimizes this recorded type cost, not token- or character-level distance. Feasibility checks schema and visible citation IDs, not entailment (Appendix A). The policy selects

$$\pi_{\mathrm{det}}(x) = \begin{cases} \arg\min_{z \in \mathcal{Z}_K : F(z,E)=1} C(o, z), & \text{if a feasible candidate exists,} \\ \perp, & \text{otherwise,} \end{cases} \tag{9}$$

with deterministic tie-breaking. The certificate is an auditable record of the operational checks and cited evidence, not proof of support or real-world truth.

This definition is for public fixed-pool comparisons. Controlled Stage3 instead adds a perturbation-family-dependent abstention penalty; its family-coded row ID also enters candidate construction and fallback seeding. Baselines do not apply the same explicit family-dependent score term, so Stage3 is a condition-aware diagnostic, not an equal-information deployment comparison. Its metadata-masked replay retains the family-coded ID and is not fully family-blind (Appendix A).

Shared-pool controls distinguish constrained selection from within-feasible-set ranking. Cost-only removes feasibility; random feasible keeps it but samples uniformly. The retained public TAFER selector also uses confidence before candidate ID to break cost ties, whereas cost-only uses ID. Their frozen contrast therefore includes tie-breaking, not filtering alone. The exact CPU replay holds cost/confidence/ID fixed to isolate filtering; the feasible cost/ID sensitivity is separate. None of these finite-protocol contrasts is a causal decomposition of model behavior.

Two conditional properties clarify the scope. First, if F accepts only supported answers on the audited relation class, every accepted candidate is supported within that class. Second, if $\mathcal{Z}_K \subseteq \mathcal{Z}_{K'}$ and the checker and cost are fixed, the minimum feasible recorded cost cannot increase as the candidate budget grows. Neither statement implies monotonic empirical accuracy under imperfect generation or checking.

Finite-pool accounting. On a fixed answer-required population, let A_K be the fraction of instances with a task-correct, support-admissible candidate under the declared evaluation relation. Let G_K be the fraction with such a candidate surviving the certificate. Let S_K be the fraction for which the selector returns a candidate that is task-correct and support-admissible under that same relation;

merely returning a non-abstaining candidate does not count as selector success. For a selector restricted to feasible candidates, $0 \le S_K \le G_K \le A_K \le 1$, giving

$$1 - S_K = \underbrace{1 - A_K}_{\text{candidate unavailability}} + \underbrace{A_K - G_K}_{\text{certificate exclusion}} + \underbrace{G_K - S_K}_{\text{selection loss}}. \tag{10}$$

This exact accounting is useful only with aligned candidate, selection, and evaluation records. It locates failure within the observed pool; it neither proves checker validity nor predicts another population. Abstention counts as zero selected-answer success here, unlike contract-correct abstention in task-policy success. The recovered citation-ID relation is neither answer entailment nor human truth.

An exploratory learned selector did not pass its pre-specified transfer criterion. Its features, results, and failed promotion are retained in the appendix; it is not part of the central method claim.

5 Experimental Design

Frozen repair comparisons. The controlled core uses Qwen2.5-7B-Instruct and seven perturbation families: clean, evidence shuffle, distractor injection, support deletion, near-miss entity substitution, relation inversion, and insufficient evidence. Its 4,200 evaluation rows are seven perturbations of 600 synthetic base examples (300 per domain); `base_id` is the repeated-measures unit, not an independently sampled population frame. In Stage3, configured TAFER applies a family-dependent abstention penalty, and the family-coded row ID enters candidate-construction and fallback seeding. The baselines do not use the same explicit family-dependent score term; we therefore treat Stage3 as a condition-aware diagnostic, not deployment evidence (Appendix A). Historical runs with disallowed hardware traces remain quarantined. Bounded public shards cover HotpotQA (Yang et al., 2018), a text-enriched FEVER-derived diagnostic slice (Thorne et al., 2018), SciFact (Wadden et al., 2020), and a separately scoped 2Wiki analysis (Ho et al., 2020). The clean Mistral comparison is retained regardless of direction.

Shared pools and observable selection. Public-data records contain six candidate methods per instance. Nine policies use the shared pools: raw, self-correction, coverage-targeted abstention, verifier-only, best-of-K, random feasible, cost-only, deterministic TAFER, and RARR-like revision (Gao et al., 2023a). The latter is a bounded pool policy, not a reproduction of the full RARR system. For these shared-pool comparisons, selection precedes the evaluation-gold join. The frozen random-feasible policy samples one feasible candidate uniformly when the feasible set is nonempty and abstains otherwise; its table reports realized per-instance outcomes, not an analytic expectation. A separate CPU replay computes the analytic uniform-feasible expectation on the same feasible set after filtering non-abstaining certificate-accepted candidates. These are distinct comparisons; subpool draws are sensitivity, not new examples. Tied generation indices do not establish nested K ordering. Frozen comparisons, matched-row accounting, and complete-pool replay have distinct populations (Appendix L); they are not a longitudinal repair test on the behavioral cohort.

Historical behavioral reanalysis. The historical condition-specific-generation ledger has 2,584 rows for 323 unique instances (2,583 PASS; one uncited-removal error). It covers full, identical-prompt sham, cited and uncited removal, cited-only, citation-ID rename, matched replacement, and no evidence. Full/sham prompts match in all 323 pairs, making sham a repeat-generation control, not an independent seed; changed-evidence prompts differ from full. Pair-specific complete cases preserve source clusters; gold joins post-generation. TextEM scores emitted text, AnsweredEM also requires a non-abstaining decision, and neither is support truth. The earlier historical generic FEVER lane lacks a validated response/abstention mapping; this limitation does not apply to the separately audited fresh V87 task-class mapping.

Bounded prospective controls. The new rule task uses 240 worlds, two instruction forms, and six evidence conditions, yielding 2,880 generation requests. Full and sham use identical prompts; anchor and irrelevant removal delete three facts each; ID renaming preserves the solver's answer set. Half the worlds have an independent alternate proof. Program truth never enters the payload. This task is separate from the historical 640-world corrected task. A structure-selected natural recovery capped HotpotQA, FEVER, and 2Wiki at 100, 80, and 100 instances, with at most two per source.

Its seven-condition Qwen-Turbo generation excludes the original answer. The initial API circuit interruption was subsequently diagnosed and the remaining frozen jobs were attempted without resending completed requests. Its results remain diagnostic-only and do not enter the primary paper comparisons.

Metrics and uncertainty. We report protocol-specific FGRs with coverage, selective FGR, task-policy success, and secondary risk summaries. Candidate-pool comparisons use task correctness with a gold citation-ID subset relation; controlled Stage3 and behavioral intervention analyses use their named operational-checker relations. These endpoints are not pooled, and neither is human-calibrated. For behavioral contrasts, the unit is a unique instance; the bootstrap samples real source clusters. Historical repair reconciliation retains 10,000 replicates and seed 17; the recovered behavioral and new rule-task analyses use 2,000 replicates and seed 20260919. These intervals are pointwise. A separate retrospective control-family sensitivity uses joint bands; complete-pool replay and that sensitivity each use 10,000 draws, with seeds 20260919 and 20260923 respectively. Equal-source weighting is a sensitivity analysis, distinct from the instance-weighted estimand. Expanded trials do not supply an independent Bernoulli sampling contract by themselves. Complete-case counts, missingness and conservative missing-outcome bounds accompany the contrasts.

Truth and provenance. Available raw generations, candidate files, checker implementations, selection receipts, source maps and hashes are indexed separately from aggregate-only artifacts. Candidate feasibility and behavioral support use different implementations in the audited lineage; this does not imply independent errors. The historical human return (Appendix Table 16) is unadjudicated and does not validate fresh outputs. Human, programmatic, task-gold and operational-proxy evidence are not substituted for one another.

6 RESULTS

Filtering and ranking are distinct contrasts. Table 1 places shared-pool controls beside TAFER. FEVER's legacy gold citation-ID-proxy FGR is below raw and cost-only selection, but sampled random feasible is nearly tied at matched coverage; HotpotQA and SciFact do not establish an incremental ranking advantage. A retrospective FEVER endpoint audit (481 instances, 105 sources; Appendix R.1) separates task-policy success from answering behavior: versus a same-key unconstrained selector with cost/confidence/ID tie-breaking held fixed, composite success rises by 0.227 $[0.189, 0.264]$, while coverage falls by 0.254 $[-0.296, -0.212]$ and correct-answer yield by 0.162 $[-0.204, -0.122]$. Versus analytic uniform-feasible selection, the composite difference is -0.002 $[-0.011, 0.003]$. Of 2,963 records, 283 match a FEVER class but have `abstain=true`; the stored answered-correct flag requires `abstain=false`. This is an endpoint-definition difference, not evidence of conflicting scorer implementations. We therefore report a legacy citation-proxy composite, not FEVER EM or human support truth. The clean Mistral comparison is adverse on gold citation-ID-proxy FGR: TAFER minus raw is $+0.0135$ $[-0.0270, 0.0631]$, a positive point estimate for risk, with an interval spanning zero. It is neither evidence of superiority nor of a universal disadvantage.

On controlled Stage3, configured TAFER $K = 4$ has coverage/task-policy success 0.5829/0.8686, versus 0.5814/0.6240 for raw and 0.6471/0.7819 for self-correction. These are condition-aware diagnostics, not equal-information deployment evidence: the configured objective applies a family-dependent abstention penalty and the family-coded ID enters candidate-construction and fallback seeding; baselines do not use the same explicit score term. Removing only the explicit family field while holding that ID fixed changes 552/4,200 selections (all abstention-to-answer), raising coverage and task-policy success each by 0.1314 (95% base-ID-cluster interval $[0.1207, 0.1426]$); it is not fully family-blind. Current replay matches configured $K = 4$ decision fields on 4,200/4,200 rows, but certificate relation labels differ on 2,450 and the historical remote source hash is unavailable (Appendix A). The constructed panel had no observed false-green events under the tested protocol (0/4,200 rows; 0/600 base IDs); all 2,448 answers passed the certificate. This is feasibility-conditioned, not human validation or a population bound.

Evidence conditions yield different observed response modes. The historical condition-specific-generation lane decomposes an operational-judge-defined supported-answer-loss proxy

Table 1: Shared-pool comparison: gold citation-ID-proxy FGR / answered coverage. The proxy requires the cited ID set to be a subset of stored gold support IDs; it is not semantic entailment or human support. The random-feasible column is the frozen policy's realized per-instance outcome: it samples one feasible candidate uniformly when available and abstains otherwise. It is not an analytic expectation. Denominator n/S is instances / source clusters, mapped from the frozen pool by unique instance ID, including abstentions.

Regime	n / S	Raw	Cost-only	Random feasible	TAFER
HotpotQA	468/78	0.1026 / 0.7628	0.1132 / 0.8675	0.0812 / 0.8333	0.0940 / 0.8333
FEVER	481/105	0.0977 / 0.5655	0.1788 / 0.7422	0.0062 / 0.4886	0.0042 / 0.4886
SciFact	202/202	0.2624 / 0.6782	0.2871 / 0.7871	0.2475 / 0.7327	0.2475 / 0.7327
Mistral HotpotQA	222/37	0.0766 / 0.7658	0.0811 / 0.8333	0.0766 / 0.7748	0.0901 / 0.7748

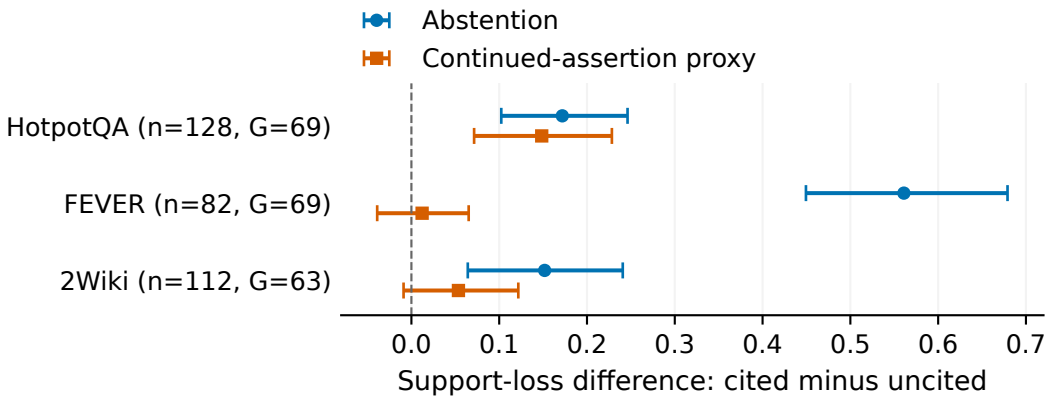


Figure 1: Historical condition-specific-generation response-mode contrasts: cited minus uncited removal. Each condition generates a response; the identical-prompt sham is a repeat-generation control, not an independent seed. Points are abstention and continued-assertion proxy components; bars are pointwise source-cluster bootstrap intervals. Complete instances / sources: HotpotQA 128/69, FEVER 82/69, 2Wiki 112/63. Both components use the full complete-triple denominator, not only initially supported answers. This historical FEVER response interface is not pooled with fresh NEI-category outputs.

into abstention and continued-assertion components (Figure 1). FEVER is abstention-dominated: $+0.561$ $[+0.449, +0.679]$ versus $+0.012$ $[-0.039, +0.065]$; HotpotQA shows both ($+0.172$ $[+0.102, +0.246]$ and $+0.148$ $[+0.071, +0.228]$), while 2Wiki's assertion estimate is uncertain ($+0.054$ $[-0.009, +0.122]$). A retrospective joint-interval check preserves the HotpotQA components and FEVER abstention against uncited-removal and sham controls, but FEVER/2Wiki assertion contrasts against sham include zero (Appendix P). These are proxies, not semantic-error or human-support labels.

A separate V87 regeneration reuses 280 instance IDs from the historical cohort; it is a new generation protocol on reused instances, not an independent-instance replication. Cited-versus-matched-uncited removal raises supported-answer-loss proxy by $+0.31$ $[+0.21, +0.41]$ in HotpotQA (100 instances/62 sources) and $+0.20$ $[+0.11, +0.30]$ in 2Wiki (100/60). Objective TextEM for uncited minus cited removal is $+0.17$ $[+0.08, +0.27]$ in HotpotQA and $+0.10$ $[-0.01, +0.22]$ in 2Wiki, whose interval includes zero. On the fixed 30-instance cross-judge subset, FEVER task-class accuracy and both judges' operational proxy contrasts align directionally; HotpotQA/2Wiki proxy signs differ and intervals span zero (Appendix N). Separately, fresh V87 FEVER class-label TextEM is 0.8375 under full/uncited removal versus 0.2625 after cited removal (80 instances/67 sources; paired difference $+0.575$ $[+0.430, +0.718]$). All outputs were non-abstentions, and scoring uses the original benchmark labels without relabeling after evidence edits. A shift to NOT ENOUGH INFO is not by itself an inappropriate answer; this is original-label task consistency, not evidence of semantic support loss or hallucination (Appendix L).

Text correctness and answering decisions differ. The objective QA comparison uses uncited removal minus cited removal. HotpotQA has 128 complete instances from 69 sources: TextEM differs by +0.109 [+0.030, +0.198] and AnsweredEM by +0.133 [+0.045, +0.224]. On 2Wiki, 112 instances from 63 sources give TextEM -0.009 [-0.071, +0.054] and AnsweredEM +0.054 [-0.018, +0.133]; both intervals contain zero. Equal-source weighting leaves these interpretations unchanged. We retain answer–abstain conflicts rather than silently converting emitted text into a committed answer. FEVER in the historical condition-specific-generation lane remains unscored; fresh V87 classification is separate, and NOT ENOUGH INFO is a task class, not refusal.

Explicit rules do not remove every response-policy error. A separate programmatic task isolates proof structure. After removing the only target proof in single-proof worlds, the explicit-rule condition returned the programmatically unsupported distractor-chain badge in 117 of 120 worlds and abstained in 3, compared with 42 distractor badges and 78 abstentions under an implicit instruction. These are answers for a different subject, not proof completions for the query. An alternate target proof remains available in redundant worlds; under explicit instructions, anchor-removal answered accuracy is 0.850, compared with 0.933 on full evidence. The implicit prompt does not state the inference rule, so its score is a task benchmark, not evidence of violating an explicitly supplied contract. Sham and ID-renaming differences are retained in the appendix rather than treated as independent seeds.

Candidate availability and ranking are distinct bottlenecks. Figure 2 summarizes frozen repair-only pools under legacy task-equality/gold-ID coverage: controlled Qwen has no candidate-level ceiling, FEVER has high feasible-good availability, and SciFact/2Wiki show candidate limits. The 2Wiki behavior sample is unlinked to its pool (Appendix Q). In a separate 2Wiki pool, a task-correct candidate satisfying the citation-ID contract appeared in 17/160 answer-required instances; TAFER selected one in 7/160. Across the full 400-instance, 80-source pool, the task-success contrast versus analytic uniform-feasible selection was -0.0038 $[-0.0134, +0.0058]$ (Appendix R.2). No unique generation order supports nested-K claims; subpool draws are not new instances.

On complete pools (Tables 24–25), with cost/confidence/ID tie-breaking fixed, filtering changes TAFER-minus-unconstrained task-policy success by $+0.0278$ $[0.000, 0.058]$ in HotpotQA (468/78 instances/sources), $+0.0594$ $[0.025, 0.099]$ in SciFact (202/202), and $+0.0495$ $[0.009, 0.095]$ in Mistral HotpotQA (222/37); these are pointwise 95% source-cluster intervals. Within the same feasible sets, confidence tie-break estimates in that order are 0.0000 $[0, 0]$, -0.0050 $[-0.0198, 0.0099]$, and 0.0000 $[0, 0]$. TAFER-minus-analytic-uniform task-policy intervals include zero in all four domains; this is inconclusive, not evidence of equivalence. In the separate FEVER citation-set analysis (481 instances/105 sources), filtering changes returned-admissible yield by $+0.0312$ $[0.0139, 0.0517]$ while coverage changes by -0.2536; TAFER versus analytic uniform-feasible yield differs by $+0.0005$ $[0, 0.0016]$, contributed by one instance in one source cluster, not a general ranking gain. These retrospective finite-pool contrasts distinguish filtering from ranking; they are descriptive, not evidence of human support validity or a general ranking advantage. Familywise intervals and source-deletion details are in Appendix F.

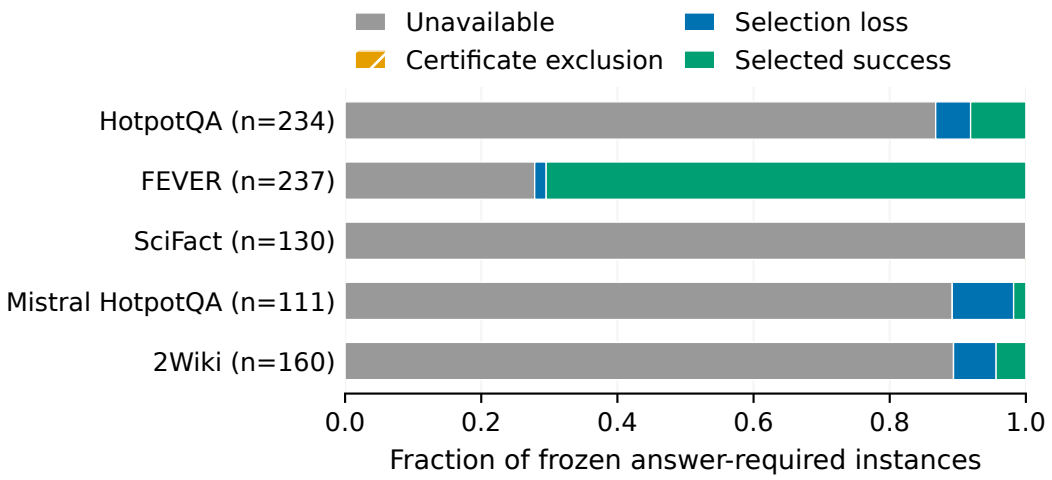


Figure 2: Frozen repair-only reference scope, independently recomputed from candidate and selection records. The three loss terms sum to one minus selected success; the losses and selected success together sum to one. Legacy task equality and gold-ID coverage define this accounting, not semantic support truth. FEVER full-pool sensitivity is reported separately; controlled Qwen has no candidate-level bar.

7 Analysis

Response mode and judge choice change the interpretation. FEVER's historical support-loss contrast is largely abstention-driven; HotpotQA has a positive continued-assertion proxy component, while the 2Wiki interval includes zero. In the fixed 30-instance sensitivity, FEVER's proxy direction is positive under both judges; HotpotQA and 2Wiki point estimates have opposite signs, but judge-specific and paired intervals include zero. These operational proxies may share errors and do not calibrate support truth. The same-prompt sham measures protocol variation, not an independent seed; renaming does not establish invariance. Historical judge–annotator comparisons also do not calibrate fresh outputs (Appendix N).

Filtering is not ranking. For a fixed pool, success can be limited by candidate absence, certificate exclusion, or selection after feasibility; this accounting does not predict another population. TAFER-minus-analytic-uniform-feasible task-policy intervals include zero in all four domains, so ranking benefit is not established. FEVER's filtering-related composite gain accompanies lower coverage and correct-answer yield; its tiny citation-yield contrast comes from one instance/source. HotpotQA remains unresolved, SciFact bounded, 2Wiki candidate-scarce, and Mistral negative; the learned selector's promotion gate failed (Appendix F).

Calibration and coverage remain necessary. Blanket abstention can yield zero checker-FGR; certificate filtering also constrains that endpoint. Neither establishes support truth, and separate selection/evaluation code does not imply independent errors. Historical author–annotator agreement is 84/150 ($\kappa = 0.3327$); selected follow-up is not sample-wide or fresh-output calibration (Appendix Table 16). Table 4 reports joint ID-rule disagreement, not conditional false-acceptance rates.

8 Limitations

Natural-language support remains an uncalibrated operational proxy; historical agreement and selected re-review do not calibrate fresh outputs. Prior AI exposure and unknown inclusion weights limit the human audit. N/A and FEVER's NEI are not abstention; the fresh 80-instance FEVER check measures original-label agreement after evidence changes, not response appropriateness. Equal-count deletion need not be semantically matched, and answer-string changes are not semantic errors. Controlled Qwen is aggregate-only; 2Wiki behavior is unlinked (Appendices K, M, and Q).

AI USE STATEMENT

Generative AI tools were used for research brainstorming and conceptual framing; proposing and refining hypotheses; feedback on methodology and experiment design; implementation and debugging of data-processing, experiment-orchestration, evaluation, statistical-analysis, and document-generation code; assistance with data cleaning and reformatting; interpretation and organization of experimental results; literature-search assistance and reference formatting; and drafting, editing, and reviewer-style critique of the manuscript and supplementary documentation. This assistance also covered mathematical formulations and explanatory derivations, and implementation of the programmatic synthetic-task generators. The synthetic diagnostics use program-enumerated proof relations, not AI-generated human labels. Model-generated statements and internal reviews are not treated as independent evidence. AI-assisted code was subjected to unit tests, dry runs, provenance and leakage audits; references have recorded primary-source checks; and paper-facing claims are linked to stored experimental artifacts. Earlier AI-preannotated and AI-assisted review materials are excluded as independent calibration. The later human annotator return was transcribed and analyzed without automated label inference or adjudication; no-AI and no-author-label conditions are author-reported, not independently verified (see limitations). The author's prior exposure to AI-assisted materials is disclosed in the audit limitations. The authors remain responsible for final verification and disclosure of the manuscript, claims, code, data processing, and reported results.

AI-assisted code also prepared blinded second- and third-pass packets and mechanically joined the returned choices to private mappings. The selected re-reviews are reported separately; no earlier label or primary agreement estimate is replaced, and no human rationale or consensus is generated by the software.

A bounded DeepSeek–Qwen rejudging analysis used frozen outputs and retained all disagreement; model ratings were not entered as human annotations.

REPRODUCIBILITY STATEMENT

The accompanying supplement provides an executable, text-free statistical reanalysis of derived FEVER outcomes for 481 unique instances in 105 source clusters. It recomputes five selector summaries and paired endpoint contrasts with 10,000 source-cluster bootstrap replicates, and checks 55 summary cells against the frozen v4 aggregate reference. This is a partial statistical reproduction only: it does not reproduce candidate generation, model inference, selector execution, natural-language intervention analyses, or the full paper. Those analyses require the underlying raw inputs and provenance described in the appendix. The released derived table contains opaque identifiers and endpoint values, but no claim/evidence text, gold strings, original identifiers, raw model responses, or private annotation material. Software tests and a successful PDF build do not independently validate the scientific claims.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=33XGfHLtZg.

Lawrence D. Brown, T. Tony Cai, and Anirban DasGupta. Interval estimation for a binomial proportion. *Statistical Science*, 16(2):101–133, 2001. doi: 10.1214/ss/1009213286. URL https://doi.org/10.1214/ss/1009213286.

Jan Buchmann, Xiao Liu, and Iryna Gurevych. Attribute or abstain: Large language models as long document assistants. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8113–8140. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.463. URL https://aclanthology.org/2024.emnlp-main.463/.

Catherine Chen, Jingyan Shen, Zhun Deng, and Lihua Lei. Conformal tail risk control for large language model alignment. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 8955–8978. PMLR, 2025. URL https://proceedings.mlr.press/v267/chen25bd.html.

Yung-Sung Chuang, Benjamin Cohen-Wang, Zejiang Shen, Zhaofeng Wu, Hu Xu, Xi Victoria Lin, James R. Glass, Shang-Wen Li, and Wen-Tau Yih. Selfcite: Self-supervised alignment for context attribution in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10839–10858. PMLR, 2025. URL https://proceedings.mlr.press/v267/chuang25a.html.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404. URL https://doi.org/10.1093/biomet/26.4.404.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. Contextcite: Attributing model generation to context. In *Advances in Neural Information Processing Systems 37*, 2024. doi: 10.52202/079017-3035. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/adbea136219b64db96a9941e4249a857-Abstract-Conference.html.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. OR-bench: An over-refusal benchmark for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 11515–11542. PMLR, 2025. URL https://proceedings.mlr.press/v267/cui25a.html.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3563–3578, 2024. doi: 10.18653/v1/2024.findings-acl.212. URL https://aclanthology.org/2024.findings-acl.212.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. RAGAS: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pp. 150–158, 2024. doi: 10.18653/v1/2024.eacl-demo.16. URL https://aclanthology.org/2024.eacl-demo.16.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16477–16508, 2023a. doi: 10.18653/v1/2023.acl-long.910. URL https://aclanthology.org/2023.acl-long.910.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488, 2023b. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Sean Welleck, Luke Wu, Jiangfeng Zhou, and Luke Zettlemoyer. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics:EMNLP 2020*, pp. 1307–1323, 2020. doi: 10.18653/v1/2020.findings-emnlp.117. URL https://aclanthology.org/2020.findings-emnlp.117.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Minlie Huang, Nan Duan, and Weizhu Chen. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Sx038qxjek.

Anna Hedström, Salim I. Amoukou, Tom Bewley, Saumitra Mishra, and Manuela Veloso. To steer or not to steer? mechanistic error reduction with abstention for language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 22924–22945. PMLR, 2025. URL https://proceedings.mlr.press/v267/hedstrom25a.html.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, 2020. doi: 10.18653/v1/2020.coling-main.580. URL https://aclanthology.org/2020.coling-main.580.

Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. Knowledge-centric hallucination detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6953–6975. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.395. URL https://aclanthology.org/2024.emnlp-main.395/.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=IkmD3fKBPQ.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval augmented generation systems. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/33dffa2e3d2ab74a783d1a8c292f66d9-Abstract-Conference.html.

Mintong Kang, Nezihe Merve Gürel, Ning Yu, Dawn Song, and Bo Li. C-rag: Certified generation risks for retrieval-augmented language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22963–23000. PMLR, 2024. URL https://proceedings.mlr.press/v235/kang24a.html.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4110–4124, 2021. doi: 10.18653/v1/2021.naacl-main.324. URL https://aclanthology.org/2021.naacl-main.324.

Minjae Lee, Kyungmin Kim, Taesoo Kim, and Sangdon Park. Selective generation for controllable language models. In *Advances in Neural Information Processing Systems 37*, pp. 50494–50527. Neural Information Processing Systems Foundation, 2024. doi: 10.52202/079017-1600. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5a6815122f533193a022cbc41786c1cc-Abstract-Conference.html.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Jiawei Li. The citation expansion trap: Operating-point control for fixed-answer multi-hop RAG citation repair. OpenReview, Grounding Language Models: Learning Faithfully and Efficiently at EMNLP 2026, 2026. URL https://openreview.net/forum?id=VYsdN4dpBJ. Accepted archival long paper.

Yifei Li, Xiang Yue, Zeyi Liao, and Huan Sun. Attributionbench: How hard is automatic attribution evaluation? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14919–14935. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.886. URL https://aclanthology.org/2024.findings-acl.886/.

Evan Z. Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6781–6792, 2021. URL https://proceedings.mlr.press/v139/liu21f.html.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, 2023. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.173. URL https://aclanthology.org/2020.acl-main.173/.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, 2023. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10862–10878, 2024. doi: 10.18653/v1/2024.acl-long.585. URL https://aclanthology.org/2024.acl-long.585.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. Automatically correcting large language models: Surveying the landscape of diverse automated correction strategies. *Transactions of the Association for Computational Linguistics*, 12:484–506, 2024. doi: 10.1162/tacl_a_00660. URL https://aclanthology.org/2024.tacl-1.27.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840, December 2023. doi: 10.1162/coli_a_00486. URL https://aclanthology.org/2023.cl-4.2/.

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–41, 2000. doi: 10.21314/JOR.2000.038. URL https://doi.org/10.21314/JOR.2000.038.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 338–354, 2024. doi: 10.18653/v1/2024.naacl-long.20. URL https://aclanthology.org/2024.naacl-long.20.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Sina Tayebati, Divake Kumar, Nastaran Darabi, Dinithi Jayasuriya, Theja Tulabandhula, Ranganath Krishnan, and Amit Ranjan Trivedi. Cap: Conformalized abstention policies for context-adaptive risk management for llms and vlms. In *Proceedings of the 17th Asian Conference on Machine Learning*, volume 304 of *Proceedings of Machine Learning Research*, pp. 926–941. PMLR, 2025. URL https://proceedings.mlr.press/v304/tayebati26a.html.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, 2018. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 7534–7550, 2020. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259.

Runtian Zhai, Chen Dan, J. Zico Kolter, and Pradeep Ravikumar. DORO: Distributional and outlier robust optimization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12345–12355, 2021. URL https://proceedings.mlr.press/v139/zhai21a.html.

A Conditional Properties and Statistical Reporting

Certificate soundness. Assume $F(z, E) = 1$ only if z is supported by E for every relation in the audited checker class. Equation 9 selects only candidates with $F = 1$; therefore every non-abstaining output is supported within that class. This is a conditional proposition, not a claim that the present structural citation/schema checker is sound for natural-language support; that soundness is not established here, and the statement does not extend to relations outside the audit.

Budget monotonicity. If $\mathcal{Z}_K \subseteq \mathcal{Z}_{K'}$ and F and C are fixed, the feasible set for K is a subset of that for K'. Hence the minimum feasible recorded cost cannot increase. This does not imply monotone empirical accuracy when candidate generation or checking changes with K.

Operational mechanism contrasts. Selectors operate on the same gold-free pool. Isolating feasibility filtering requires both recorded-cost ranking and tie-breaking to remain fixed; the retained public TAFER/cost-only comparison differs in confidence tie-breaking and does not meet that stricter condition. Comparing with random feasible instead measures the entire within-feasible-set selection policy. Neither contrast is a causal decomposition of the generator or checker.

Recovered operational implementation. The retained public generator assigns raw candidates cost 0, independently generated alternatives 0.5, and self-correction/RARR-like candidates 1. These type-assigned costs are present in all four frozen main-table pools; they are not Levenshtein distances or measured editing effort. The retained checker accepts a schema-valid answer with nonempty citation IDs contained in the visible-ID set (or a schema-valid abstention); selection additionally requires a non-abstaining feasible answer, denoted by $F = 1$. Its binary verifier score records visible-ID validity, not entailment strength. Behavioral support is judged separately as V; the post-selection citation-set evaluation is R^{ID} (Appendix Q). Source-code inspection and pooled field consistency document these operational limits, but do not independently authenticate every historical executable version. A cost/ID CPU replay is not identical to the frozen cost/confidence/ID selector and must retain its separate scope.

Controlled Stage3 information boundary and selector. The Stage3 wrapper passes the full perturbation row to selection. In the audited local source, generation uses the question and visible evidence text; the output parser filters citations against visible evidence IDs. Repair candidate construction reads the question, evidence IDs/text, and row ID; the row ID has the form `base_id::perturbation` and seeds candidate construction. The TAFER objective additionally reads the explicit perturbation-family field. The local selector does not directly read `world_answer`, `gold_answer`, `expected_action`, `support_ids`, or `supported_answers`, although those fields are present in the full row passed by the wrapper. This code-path audit is not proof of the historical remote execution boundary: the remote source hash is unavailable.

For TAFER, the controlled Stage3 score is recorded candidate cost, plus 100 for an infeasible answer, plus 0.35 for abstention, plus 0.40 for abstention when visible candidates exist and the explicit family is not support deletion, near-miss entity, relation inversion, or insufficient pool, and minus 0.15 for a feasible answer. In the local code, no-repair, self-correction, and add-only do not branch on the explicit perturbation field; TAFER does. However, self-correction's random branch and TAFER's candidate-pool fallback are seeded from the row ID, whose suffix encodes the family. The comparison therefore retains unequal semantic side information and a family-coded seed channel; it is neither an equal-information deployment comparison nor a clean family-access ablation. This objective is distinct from the public fixed-pool minimum-recorded-cost selector.

A current-source CPU replay has matching row keys for all 29,400 stored repair rows; 15,923 match in every field and 13,477 do not. For configured TAFER $K = 4$, answer, abstention, confidence, cited IDs, repair cost, certificate-pass decision, and output text match on 4,200/4,200 rows. Certificate relation metadata differs on 2,450 rows, so full-row exact replay does not pass. The older 29,400-row exact-replay claim is withdrawn. The explicit-family-field ablation below holds the family-coded row ID fixed; it is not a fully family-blind replay. These are local reconciliations, not cryptographic identification of the historical remote source revision.

The explicit-family-field ablation reuses the same stored generations and evaluation rows, removing only the explicit family field while retaining the family-coded row ID and therefore the candidate-construction seed. It is a retrospective CPU sensitivity, not a new model run or a replacement primary analysis. Its `base_id` bootstrap treats the seven perturbation variants of each synthetic example as one repeated-measures cluster; the synthetic examples are not asserted to be a probability sample. Table 2 gives the exact finite-cohort rates and selection transitions for this narrow ablation.

For a zero-event sensitivity, a base ID is event-positive if any of its seven perturbation rows records a false-green event. The observed count is $0/600$. If these base IDs are treated as independent Bernoulli units, the one-sided exact 95% Clopper–Pearson upper limit is $1 - 0.05^{1/600} = 0.00498$ (0.50%). The synthetic panel is not a probability sample, so this assumption-based limit is not a population-risk bound. The expanded $0/4{,}200$ row count is not used as an independent-trial denominator. Table 2 also reports $0/4{,}200$ for both selectors; every observed non-abstention passed

Table 2: Retrospective explicit-family-field ablation on the same 4,200 stored generations, retaining the family-coded row ID and candidate seed. This is not a fully family-blind replay or a new model run. Task-success and coverage differences are each $+0.1314$ (95% `base_id`-cluster bootstrap interval $[+0.1207, +0.1426]$; 10,000 draws, seed 17), not a population guarantee.

Selector	Task success	Coverage	FGR events	Changed
Configured family-conditioned	0.8686	0.5829	0/4,200	–
Explicit-family-field ablation	1.0000	0.7143	0/4,200	552/4,200
Ablation minus configured	$+0.1314$	$+0.1314$	0	–
Near-miss entity	–	–	–	252/600
Relation inversion	–	–	–	300/600
Other five families	–	–	–	0/3,000

Table 3: Operational contract used by the recovered candidate-pool analysis.

Component	Frozen contract
Generation input	`instance_id`, prompt, group, and source ID only; gold and evaluation-only fields are joined after selection.
Structural checks	Schema-valid output with nonempty citation IDs contained in the visible-ID set, or a schema-valid abstention.
Certificate meaning	A structural citation/schema check; it is not entailment, human support, or task truth.
Selector	Lowest recorded type cost among non-abstaining feasible candidates; the retained frozen selector uses confidence and candidate ID for cost ties.
Cost scale	Type-assigned values: raw 0, independently generated 0.5, and self-correction/RARR-like 1; these are not measured edit distances.
Evaluation boundary	Task correctness and support fields are evaluation-only joins; operational-proxy, legacy-contract, and human routes are reported separately.

the structural certificate (2,448 configured; 3,000 visible-only). These are feasibility-conditioned observations in this controlled panel, not independent support validation or population estimates.

An artifact-specific audit of the frozen candidate pools and V82 behavioral-support ledger found zero prompt-hash intersection: F is deterministic schema/visible-ID code, while V is a separately prompted Qwen-Turbo `strict_claim_entailment` judge. This rules out literal same-checker reuse only for those audited artifacts; it does not establish independent errors or human validity.

For reproducibility, the operative contract is summarized in Table 3. The table is a protocol description, not an additional validity result: the certificate is deliberately weaker than semantic support, and the returned role labels remain an unresolved calibration diagnostic.

Two stored-ID rules, not semantic calibration. This historical diagnostic uses answer-required, non-abstained rows with complete identifier fields. The generator's `evidence_ids` field contains returned citation IDs Q, not the visible-context inventory. With stored gold IDs G, the legacy support field is $P = \mathbb{1}[Q \neq \varnothing, Q \subseteq G]$, whereas the comparison rule is $R = \mathbb{1}[G \neq \varnothing, G \subseteq Q]$. Agreement is 0.495 (216 rows, 76 sources) on Qwen HotpotQA, 0.911 (203, 64) on FEVER, and 0.560 (91, 34) on Mistral HotpotQA. The former "false accept" values 0.472, 0.089, and 0.308 are *joint* $P = 1, R = 0$ proportions over all complete rows, not $\Pr(R = 0 \mid P = 1)$. The historical numbers and source-cluster intervals remain unchanged in Table 4; neither predicate establishes answer entailment, complete semantic support, or human truth. Their disagreement is between ID inclusion rules, not a measurement of independent semantic-checker error.

Synthetic contract examples (not data). The boundary above is concrete even when a record is structurally valid. The following toy cases are explanations of the frozen predicates, not additional experimental observations.

Candidate availability, certificate survival, and accounting. For instance i, let $\mathcal{Z}_{K,i}$ be the generated candidate pool and let $\mathcal{F}_{K,i} = \{z \in \mathcal{Z}_{K,i} : F(z, E_i) = 1\}$ be its feasible subset. Let

Table 4: Two ID-rule diagnostic on complete answered rows. Every displayed proportion uses n; intervals retain the original pointwise source-cluster bootstrap. Neither rule is semantic support truth. A degenerate zero interval is descriptive, not a risk bound.

Regime	n	clusters	Agreement	$P{=}1, R{=}0$	$P{=}0, R{=}1$
Qwen HotpotQA	216	76	.495 [.407,.584]	.472 [.382,.566]	.032 [.009,.062]
FEVER	203	64	.911 [.833,.976]	.089 [.024,.167]	.000 [.000,.000]
Mistral HotpotQA	91	34	.560 [.432,.688]	.308 [.179,.442]	.132 [.056,.223]

Table 5: Synthetic examples separating structural checks, selection and citation-set rules.

Case	Visible IDs and output	Structural check	Citation-set interpretation
Structural pass	IDs {E1,E2}; answer cites E1	Feasible: non-abstaining, valid citation ID	Citation-subset check fails if gold IDs are {E2}
Abstention	IDs {E1}; schema-valid abstention	Schema-valid; $F = 0$	No supported-answer claim is established

$Q_i(z) = 1$ when candidate z is task-correct and supported under the evaluation-only relation. We distinguish exact-good availability from certificate survival:

$$A_K = \frac{1}{n}\sum_i \mathbb{1}[\exists z \in \mathcal{Z}_{K,i} : Q_i(z) = 1], \qquad G_K = \frac{1}{n}\sum_i \mathbb{1}[\exists z \in \mathcal{F}_{K,i} : Q_i(z) = 1]. \tag{11}$$

A_K is the finite-pool candidate ceiling and G_K is the part of that ceiling surviving the certificate. Candidate-unavailability loss is $1 - A_K$, and verifier-exclusion loss is $A_K - G_K$. Let S_K be the rate at which the deployed selector returns a task-correct, support-admissible candidate (with abstention assigned value zero). If the selector chooses only from $\mathcal{F}_{K,i}$, its finite-pool ranking regret is $G_K - S_K$. Let $t_{\pi,i}$ indicate task-policy success: it is one for a correct supported answer or a contract-correct abstention, and zero otherwise, and let $T_K = n^{-1}\sum_i t_{\pi,i}$. A false certificate acceptance is audited separately and is not treated as support truth. Define

$$s_{\pi,i} = \mathbb{1}[\pi \text{ returns a non-abstaining } z \text{ with } Q_i(z) = 1], \qquad S_K = \frac{1}{n}\sum_i s_{\pi,i}. \tag{12}$$

Here abstention has $s_{\pi,i} = 0$, so T_K may exceed S_K when abstention is the contract-correct action. Consequently, the finite-sample accounting identity is

$$1 - S_K = (1 - A_K) + (A_K - G_K) + (G_K - S_K). \tag{13}$$

This is a bookkeeping identity under the evaluation contract, not a causal decomposition or a population repairability law. Separately, for a stated binary check J and a valid support reference, conditional false acceptance is defined below. J must identify the actual check (F for selection feasibility or V for a behavioral support judge), not an interchangeable surrogate:

$$\mathrm{VFA}_K = \Pr\big[S(z, E, G) = 0 \mid J(z, E) = 1\big], \qquad \widehat{\mathrm{VFA}}_K = \frac{\sum_{i,z}\mathbb{1}[J_{i,z} = 1, S_{i,z} = 0]}{\sum_{i,z}\mathbb{1}[J_{i,z} = 1]}. \tag{14}$$

The denominator is the number accepted by J, not all rows; when it is zero the conditional rate is undefined. This definition does not relabel the joint ID-rule proportions above as semantic-checker calibration.

Conditional context sufficiency and necessity estimands. For the same source instance, let Y_i^{full}, Y_i^{selected}, and Y_i^{removed} indicate supported correctness under a declared support reference with full evidence, selected-only evidence, and the selected evidence removed. Define

$$\Delta_{\mathrm{suff}} = \mathbb{E}[Y^{\mathrm{selected}} - Y^{\mathrm{full}}], \qquad \Delta_{\mathrm{nec}} = \mathbb{E}[Y^{\mathrm{removed}} - Y^{\mathrm{full}}]. \tag{15}$$

These definitions do not constitute a completed exact-support test on the natural-language populations. The reported behavioral contrasts use task gold or an operational checker, with their own pair-specific denominators. Substituting a checker supplies a proxy contrast, not exact supported correctness. No natural-language sufficiency or necessity guarantee follows, and neither contrast identifies a causal effect outside the tested prompts.

Ranking contribution and repair cost. Within a feasible pool, let π_{det} select the lowest recorded-cost candidate and let π_{rand} sample from the same feasible set. For any finite-protocol metric M, the ranking contrast is

$$\Delta_M^{\mathrm{rank}} = M(\pi_{\mathrm{det}}) - M(\pi_{\mathrm{rand}}). \tag{16}$$

For risk metrics lower is better; for coverage and task-policy success higher is better. We call this a ranking contrast rather than a causal effect. A corresponding evaluation-only regret against the best observed feasible candidate is

$$\mathrm{Reg}_M^{\mathrm{rank}}(i) = \ell_M(z_{\mathrm{det},i}) - \min_{z \in \mathcal{F}_{K,i}} \ell_M(z), \tag{17}$$

where ℓ_M is a per-instance loss (the negative utility when M is a utility), and the minimum is computed only after gold metadata is joined; it is not available to the deployed selector. Mean repair cost is reported as

$$\overline{C}(\pi) = \frac{1}{n} \sum_{i=1}^{n} C(o_i, \pi(x_i)) \mathbb{1}[\pi(x_i) \neq \bot]. \tag{18}$$

Empirical repairability frontier. Each tested policy is represented by $\big(R_{\mathrm{FG}}, -\mathrm{Coverage}, -T_K, \overline{C}\big)$, where smaller coordinates are preferred. A policy is dominated if another policy is no worse on all four coordinates and strictly better on at least one. The empirical repairability frontier is the set of nondominated tested policies. This frontier summarizes finite-pool trade-offs among risk, coverage, task-policy success, and repair cost; it is not a learned predictor and does not establish a population frontier or a causal mechanism.

Bounded consequence loss and worst-group risk. For the finite-protocol score $T(x)$ and its empirical quantile $q_\alpha(T)$, tail FGR is $\Pr[L_{\mathrm{FG}} = 1 \mid T(x) \geq q_\alpha(T)]$. We also retain the share of bounded loss in the highest-risk 10% of instances. The frozen controlled tables use their recorded `weighted_loss` definition; absolute tail-loss values are not ranked across different scoring protocols. Each finite evaluation pipeline records a prespecified bounded consequence loss $L^{(r)} \in [0, L_{\max}^{(r)}]$. For the public-data scorer, the recorded loss is

$$L^{(\mathrm{pub})} = 5\,FG + 4\,U + O + C, \tag{19}$$

where $U = \mathbb{I}[\text{wrong} \wedge \text{unsupported} \wedge \text{not abstained}]$, $O = \mathbb{I}[\text{unnecessary abstention}]$, and C is the recorded bounded repair cost. The frozen Stage3 tables retain the inherited `weighted_loss` receipt and its sensitivity analysis rather than being silently recomputed with the public scorer. For perturbation groups $g \in \mathcal{G}$, worst-group FGR is

$$R_{\mathrm{WG}} = \max_{g \in \mathcal{G}} \frac{1}{n_g} \sum_{i: g_i = g} FG_i. \tag{20}$$

For empirical losses $L_1^{(r)}, \ldots, L_n^{(r)}$, the legacy threshold-tail statistic averages all losses at or above the empirical α-quantile, including all ties. Its historical field name is `cvar_loss`. This is not generally the empirical expected shortfall: the latter averages exactly the upper $1-\alpha$ probability mass, using fractional boundary mass when needed. We preserve the frozen values and label the controlled table as threshold-tail mean rather than silently recomputing it. Later fractional-tail implementations must be identified separately. Neither statistic, at $\alpha \in \{0.80, 0.90, 0.95\}$, establishes a population tail law.

Blanket-abstention degeneracy. For any sample, a policy that always emits $\bot$ has unconditional and selective FGR equal to zero (with selective FGR defined as zero at zero coverage). Thus FGR alone is not a sufficient objective. We jointly report coverage, task-policy success, over-abstention, and AURC, and treat the zero-coverage verifier-only policy as a degeneracy check rather than a successful repair method.

Zero-event bound. For n independent Bernoulli units with zero observed events, the one-sided upper limit is $1 - 0.05^{1/n}$. This conditional formula is not justified merely by counting expanded perturbation rows. For source-level any-event indicators it bounds a source event probability, not an output-level failure probability.

Table 6: Frozen bounded comparisons, retained without rerunning or selecting favorable outcomes. Public differences are TAFER minus raw FGR with source-cluster intervals. Support is operational-checker defined. The controlled row reports 4,200 expanded records from 600 base examples; source-level independence and a population sampling frame are unverified, so no population bound is reported.

Regime	Expanded/base n	Raw FGR	TAFER FGR	Δ FGR [95% interval]
Controlled Qwen	4,200/600	.1207	0 obs.	not pooled
HotpotQA	468	.1026	.0940	$-.0085[-.0385, .0192]$
FEVER	481	.0977	.0042	$-.0936[-.1282, -.0611]$
SciFact	202	.2624	.2475	$-.0149[-.0545, .0198]$
Mistral HotpotQA	222	.0766	.0901	$+.0135[-.0270, .0631]$

B Frozen Repair Comparisons

These are the retained historical repair estimates, not new behavioral measurements. The historical HotpotQA and FEVER bootstrap receipts record source-field fallbacks; they must not be described as newly verified explicit source maps. The behavioral reanalysis instead uses the recovered planned-job source map and reports its own denominators. No expanded-trial population bound is inferred for the controlled row.

C Full HotpotQA Baseline Table

Method	Accuracy	FGR	Coverage	Sel. FGR	AURC
Raw	.2415	.1026	.7628	.1345	.1090
Self-correction	.2415	.0983	.6816	.1442	.0906
Matched abstention	.2415	.1026	.7628	.1345	.1090
Verifier-only	.0000	.0000	.0000	.0000	.0000
Best-of-K	.2436	.1047	.8611	.1216	.1010
Random feasible	.2308	.0812	.8333	.0974	.0737
Cost-only	.2564	.1132	.8675	.1305	.1095
TAFER	.2329	.0940	.8333	.1128	.0918
RARR-like	.2201	.0812	.6239	.1301	.0807

Table 7: All nine HotpotQA selectors. Verifier-only is a zero-coverage sanity baseline.

D Full SciFact Baseline Table

Method	FGR	Coverage	Sel. FGR	Policy succ.	AURC
Raw	.2624	.6782	.3869	.4703	.3649
Self-correction	.2475	.6188	.4000	.5099	.4109
Matched abstention	.2624	.6782	.3869	.4703	.3649
Verifier-only	.0000	.0000	.0000	.3564	.0000
Best-of-K	.2723	.7871	.3459	.4554	.3561
Random feasible	.2475	.7327	.3378	.5050	.3650
Cost-only	.2871	.7871	.3648	.4356	.3880
TAFER	.2475	.7327	.3378	.5050	.3417
RARR-like	.2030	.4950	.4100	.4901	.3546

Table 8: All nine selectors on the 202-instance SciFact evaluation split. The zero-coverage verifier-only row is a degeneracy check. One malformed source generation is explicitly scored as an invalid output.

E Learned Selector Audit

The learned feature vector contains 11 inference-time observable candidate fields and excludes gold labels, oracle actions, perturbation receipts, support identifiers, and severity. ERM, GroupDRO, CVaR, and hybrid models use seeds 17, 23, and 41. Current HotpotQA source-group and held-out-family results are stored as appendix evidence. Zero-retraining FEVER transfer is complete, but no variant passes the strict two-of-three-seeds improvement gate against deterministic TAFER. Cross-architecture transfer remains unavailable. The selector is not promoted.

F Baseline Fairness Disposition

All nine public-data methods select from the same dataset-specific candidate pool, and evaluation metadata is joined only after selection. FEVER coverage-targeted abstention is within 0.0167 of TAFER coverage. On HotpotQA, TAFER coverage is 0.8333 while raw generation covers 0.7628; abstaining from raw outputs cannot increase coverage to the TAFER operating point. We therefore report the maximal raw operating point and compare complete risk–coverage curves rather than calling the two points matched.

Row-level pool membership and paired selectors. A separate receipt-locked CPU audit checked four exact-contract candidate/selection artifact pairs, each with six candidates per instance. Across 36 regime–method checks, all 8,462 non-abstaining selected rows joined to a candidate in the declared pool, with no candidate-ID, checked-field, or source-ID mismatches; SciFact is unavailable at row level. TAFER and random-feasible have identical observed coverage on their paired answer-required cohorts. Table 9 reports paired source-cluster contrasts. FEVER task-policy and answer-correctness outcomes are distinct endpoints; the separate audit reports same-key, feasible cost/ID, and analytic uniform-feasible contrasts, but does not estimate the sampled random-feasible contrast in Table 9. The composite task-policy outcome includes a citation-ID proxy rather than human support truth. The other support and certificate outcomes use an explicit evaluation-only ID contract, not human support truth. These post-hoc pointwise sensitivities verify pool lineage; they do not establish ranking superiority. Realized repair costs are measured, not equalized.

Multiplicity-aware same-pool sensitivity. We computed single-step max-$|t|$ source-cluster bootstrap intervals for the fixed family of four comparators (raw, best-of-K, random-feasible, and cost-only), four metrics (coverage, stored legacy task-success, explicit-contract supported-correct, and certificate-compliant success), and four row-level regimes. Twelve FEVER task-dependent cells are withheld because its label/answer/abstention mapping is unvalidated, leaving 52 reported cells. Eight of the 52 pointwise 95% intervals exclude zero, whereas none of the 52 simultaneous 95% intervals does. This post-hoc correction reinforces that the same-pool audit does not establish a general selector-ranking benefit. This family does not include the separate FEVER citation-set-inclusion endpoint, which remains restricted to its own operational contract and is not task accuracy or human support validity. The complete intervals, denominators, and receipt are supplied with the audit artifacts.

Single-source influence sensitivity. A leave-one-source-cluster-out analysis changes the sign of 3 of 39 nonzero- baseline endpoints; all are near-zero legacy task-success contrasts. For Qwen HotpotQA, TAFER minus raw is -0.009 (leave-one-out range $[-0.017, +0.004]$; 78 sources), and TAFER minus random-feasible is $+0.004$ ($[-0.004, +0.013]$; 78 sources). For Qwen 2Wiki, TAFER minus random-feasible is $+0.006$ ($[-0.006, +0.013]$; 80 sources). Thirteen other reported endpoints have an exactly zero full-sample contrast, so their direction stability is undefined. These are influence ranges, not confidence intervals or tests; the support and certificate outcomes remain evaluation-only contract proxies.

Table 9: TAFER minus random-feasible on fixed six-candidate pools. Entries are pointwise 95% source-cluster percentile intervals. Support and certificate outcomes are evaluation-only explicit-contract proxies. FEVER task-policy and answer-correctness endpoints for this sampled random-feasible contrast are not estimated; separate comparator-specific endpoints are reported in Appendix R.1.

Regime	Task Δ	Support Δ	Cert. Δ	n/cl.	n/cl. cert.
Qwen HotpotQA	$+0.004$ $[-0.030, +0.043]$	$+0.021$ $[0.000, +0.043]$	$+0.023$ $[0.000, +0.048]$	234/78	216/76
Qwen FEVER	withheld	withheld	withheld	237/71	203/64
Mistral HotpotQA	0.000 $[-0.063, +0.063]$	-0.018 $[-0.072, +0.036]$	-0.022 $[-0.092, +0.046]$	111/37	91/34
Qwen 2Wiki	$+0.006$ $[-0.038, +0.056]$	-0.019 $[-0.050, +0.006]$	-0.024 $[-0.062, +0.008]$	160/80	126/65

G Finite-Sample Risk Sensitivity

For HotpotQA, FEVER, and SciFact we retain within-protocol tail-loss summaries at $\alpha \in \{0.80, 0.90, 0.95\}$ and worst-group FGR after requiring at least 20, 50, or 100 examples per eligible group. Legacy threshold-tail and fractional-tail definitions must remain distinguished as above. SciFact has only one clean group, so its worst-group value is identical to its mean and is not a tail-concentration test. The machine-readable receipt remains in local audit records and is not included in the current submission files. These checks describe bounded-shard sensitivity, not a population tail law.

H Paired Finite-Sample Inference

The source-cluster paired-bootstrap receipt reports target and reference point estimates, 10,000-replicate percentile intervals, and the pairing contract for every public-data comparison, including the two mechanism contrasts against cost-only and random feasible. A separate row-level same-coverage replay applies the same 10,000-draw source-cluster procedure to TAFER versus random feasible on the four retained pools; all four FGR intervals include zero. We label intervals containing zero as not distinguishable at the current shard size and report no bootstrap p-values.

As a cluster-influence sensitivity check, we also recompute the same-coverage FGR contrast after leaving out each source cluster in turn. The direction is positive for all 78 non-zero HotpotQA exclusions, non-positive for all 104 non-zero FEVER exclusions, non-negative for Mistral with one zero exclusion, and near zero with both signs for SciFact. This reduces concern that the mixed directions are caused by one cluster, but remains an operational-proxy audit and does not establish a universal ranking benefit or equivalence.

Equal-source weighting sensitivity. The exact-selector reconciliation also reports a source-weighted sensitivity: each source contributes one mean contrast before source means are averaged, rather than weighting every instance equally. For the feasibility-filter contrast in supportedness, the source-weighted estimates are $+0.0059$ $[-0.0419, +0.0559]$ for HotpotQA, $+0.0729$ $[+0.0303, +0.1238]$ for FEVER, $+0.0338$ $[-0.0068, +0.0811]$ for SciFact, and $+0.0289$ $[-0.0544, +0.1025]$ for clean Mistral. The cost–ID tie-break contrast is zero for FEVER and Mistral, -0.0022 $[-0.0067, 0.0000]$ for HotpotQA, and $+0.0074$ $[0.0000, +0.0222]$ for SciFact. Thus source weighting changes magnitudes but not the bounded, heterogeneous interpretation; it does not establish equivalence, a ranking law, or human support validity. The receipt uses 10,000 source-cluster draws and retains the same incomplete truth-join scope.

I Nearest-Neighbor Positioning

Table 10 compares stated roles rather than claiming that prior work never discusses an unmarked feature. TAFER's distinction is the joint evaluation contract, not novelty of every component in isolation.

Work	Central evaluation role and scope
Contrast Sets (Gardner et al., 2020)	Dataset tests of local decision boundaries through contrast interventions.
RARR (Gao et al., 2023a)	Retrieval-backed revision of unsupported claims on public tasks.
CRITIC (Gou et al., 2024)	Tool-feedback critique and revision of output errors on public tasks.
SelfCheckGPT (Manakul et al., 2023)	Sampling-based detection of sentence hallucinations, with human evaluation.
Selective Generation (Lee et al., 2024)	Selective generation decisions and risk control on public tasks.
ARES (Saad-Falcon et al., 2024)	Evaluation of retrieval-augmented generation quality, with human evaluation.
ContextCite (Cohen-Wang et al., 2024)	Attribution through context perturbation on public tasks.
AttributionBench (Li et al., 2024)	Auditing attribution-evaluator reliability against human judgments.
Attribute or Abstain (Buchmann et al., 2024)	Attributed answers and abstention, with human evaluation.
RefChecker (Hu et al., 2024)	Claim-level support checking, with human evaluation.
OR-Bench (Cui et al., 2025)	Refusal decisions under over-refusal stress tests, with human evaluation.
TAFER	False-green answers, bounded interventions and certificate-constrained repair; operational support proxy, coverage and finite-pool accounting, without completed independent human calibration.

Table 10: Scoped positioning of the cited works. The table identifies central evaluation roles, not every capability; an omitted feature is not an absence claim.

J Data and Provenance Boundaries

The final model-input schema contains only `instance_id`, `prompt`, `group`, and `source_id`. Evaluation files retain gold labels and support relations but are never passed to generation, repair generation, learned candidate selection, or hidden-state extraction. Each accepted run records the resolved config, prompt hash, raw output, candidates, certificates, checker results, metrics, self-verification, node, and GPU. Runs on disallowed hardware remain excluded under the frozen provenance protocol.

Checker-lineage audit. For the four recovered candidate artifacts (Qwen HotpotQA, FEVER, SciFact, and clean-negative Mistral HotpotQA), candidate feasibility is determined by the recorded certificate and abstention fields together with visible-citation checks. The later support proxy is a separate Qwen-Turbo `strict_claim_entailment` judge ledger (2,583 rows: 2,582 passes and one error). The audited components have different implementations, and their recorded prompt-hash sets do not overlap. This rules out reuse of the identical checker implementation in this audited path; it does not establish independent errors, unbiasedness, or human support validity. The support outcomes remain operational proxies, and 2Wiki was not included in this lineage audit. The separate runtime relationship receipt is marked `SKIP_CHECKER_MANIFEST_MISSING`; therefore, this is a source-level and retained-hash distinction, not a full runtime-configuration equivalence audit.

K Historical Behavioral and Programmatic Controls

The historical behavior lane was run after generation and used a post-generation gold join. The model prompt never contained gold answers, labels, or support identifiers. The eight condition names are protocol identifiers only: full, sham repeat, citation-ID rename, uncited removal, cited removal,

matched replacement, cited-only, and no evidence. Pair-specific complete cases are reported because some conditions have different eligibility; missing rows are not imputed.

The earlier base-ID objective task table is superseded by the family-preserving reanalysis below. Its original report remains in the private audit history, but its overwritten source-condition records are not final paper evidence. The historical base-ID replay's FEVER task EM remains withheld; the separate fresh V87 FEVER mapping is documented in Appendix L and does not retroactively repair that historical replay.

The corrected RuleWorld v2 worlds use three anchor proof facts, an independent alternate proof in redundant worlds, and exactly three proof-shaped distractors. The lane contains 640 worlds across four strata and 4,480 jobs. The v1 stale-proof-ID and mismatched-removal defects are retained in the audit history but are excluded from the v2 aggregate. Redundant worlds answer correctly after anchor removal at rate 0.99375; nonredundant worlds correctly abstain at rate 0, with overcommit rate 1.0. Citation-ID rename preserves answerability on all 640 worlds. This is a programmatic contract control only.

L Recovered Measurement and Scope Audit

The historical generation ledger contains 2,584 planned rows; 2,583 have a valid generation decision. The audit retains 602 rows with an abstention decision and substantive answer text. TextEM scores the text regardless of that decision; AnsweredEM multiplies TextEM by the non-abstention indicator. Neither the answer field nor the decision is edited. An earlier task evaluator collapsed family-qualified instance IDs to a base ID and overwrote source-condition records. Its 69-instance HotpotQA and 63-instance 2Wiki task tables are superseded, not interpreted as an alternative source-weighted estimate. The present report retains full instance IDs and reports equal-source sensitivity separately.

For a complete full/cited/uncited triple, let B^t indicate abstention and let V^t be the support proxy for a non-abstaining answer. Define

$$L^t = \mathbb{1}[B^0 = 0, V^0 = 1] \left(\mathbb{1}[B^t = 1] + \mathbb{1}[B^t = 0, V^t = 0]\right).$$

The cited-minus-uncited loss decomposes pointwise into abstention and continued-assertion components. All components use the same complete-triple denominator, including zero contributions from initially unsupported or abstaining responses. Missing required support is unknown, not false. The historical HotpotQA triple population has 129 planned instances and 128 complete triples over 69 sources; FEVER has 82 over 69, and 2Wiki has 112 over 63. Inference is source-cluster based and conditional on these finite populations.

Candidate replay uses two different populations, never silently pooled. Frozen-reference reconciliation requires a matched historical TAFER selected row and preserves the original support-field precedence; all five recovered regimes match the old row-level reference. The separate observable-selector sensitivity includes every complete answer-required candidate pool, so its denominator can be larger. In FEVER these are 237 and 294 instances, respectively. Repair-only excludes raw candidates; full-pool sensitivity includes them. Here support means nonempty stored gold IDs G are covered by returned citation IDs Q, $G \subseteq Q$, for an answer-required, non-abstaining output. This is not semantic entailment or a complete support contract. The citation-subset rule $Q \subseteq G$ in Appendix Q is a different metric on different pools. The single FEVER raw-candidate discrepancy is retained under both scopes. No ordering is inferred when different candidate methods share a generation index.

Raw field and denominator verification. The retained post-generation scorer reproduces the task and ID-coverage fields of all 12,900 candidate rows in these five artifacts, and the atlas's 872 complete-instance decompositions match the frozen reference row by row. Among answer-required instances, 66 HotpotQA, 57 FEVER, 28 SciFact, and 39 Mistral records lack a matched historical TAFER selection; 2Wiki has none. They are excluded from the atlas, not assigned unsuccessful repairs. The remaining populations in Table 11 are therefore conditional complete cases, not estimates for every generated instance. Abstention placeholders without source IDs are linked through the unique candidate instance-to-source mapping, never by using instance ID as source. Legacy task correctness is normalized equality of the recorded answer or label fields; this replay does not validate an official FEVER response mapping. No non-abstaining row in these files has both prediction and target missing. The certificate-only and full structural-feasibility ceilings agree on these complete

Table 11: Distinct analysis populations. Each entry is instances/source clusters. The frozen repair main table has a further, different denominator and is not replaced by either replay. The rightmost column is the wider answer-required gold-coverage population, not the citation-subset replay in Appendix Q; neither adds independent data.

Regime	Matched-row atlas	All answer-required pools
Qwen HotpotQA	234/78	300/100
Qwen FEVER	237/71	294/88
Qwen SciFact	130/130	158/158
Mistral HotpotQA	111/37	150/50
Qwen 2Wiki	160/80	160/80

good-candidate sets, although their flags need not agree on other candidates. None of these software identities certifies semantic support.

Exploratory paired selector check. A post-review CPU check pairs the retained complete-pool selector rows, without changing their choices or support fields. For repair-only pools, recorded-cost selector minus random feasible has selected supported-correct success differences of -0.0038 for HotpotQA (pointwise 95% interval $[-0.0214, 0.0135]$), -0.0026 for FEVER ($[-0.0166, 0.0070]$), -0.0271 for Mistral ($[-0.0588, 0.0046]$), and -0.0031 for 2Wiki ($[-0.0244, 0.0184]$). SciFact has a degenerate zero difference. The check resamples source clusters 2,000 times with seed 20260919. Full-pool sensitivity gives the same differences for this endpoint. Recorded-cost and maximum-score selection also have equal observed success metrics; this need not mean they select identical candidates. These exploratory, pointwise intervals establish neither equivalence nor a positive ranking advantage and do not replace the frozen main comparisons.

Exact shared-pool selector reconciliation. We separately replayed the frozen implementation on the four paper-reference candidate pools: Qwen HotpotQA (468 instances, 78 source clusters), Qwen FEVER (481, 105), Qwen SciFact (202, 202), and clean Mistral HotpotQA (222, 37). A separately scoped prospective Qwen 2Wiki artifact is also available under a named, hashed lineage: 2,400 shared-pool rows, 400 instances, and 80 source clusters. Its stored TAFER and cost-only selections reproduce the explicit CPU keys on all 400 instances. We retain this result as evaluation-only sensitivity rather than merging it into the frozen paper-reference population; it does not establish human support truth or a uniform ranking advantage. The audit compared four selectors: the recorded TAFER key on feasible candidates, the same cost–confidence–ID key without feasibility filtering, a feasible cost–ID key, and the unconstrained cost–ID key. Gold fields were joined only after selection. The exact TAFER and cost-only choices matched the stored selector rows on every row for which a stored candidate truth record existed, but coverage was incomplete (TAFER truth-join coverage was 390/468, 235/481, 148/202, and 172/222, respectively). For the named 2Wiki artifact, TAFER minus the unconstrained key gives $+0.0071$ $[-0.0216, +0.0374]$ in supportedness and -0.0355 $[-0.0727, 0.0000]$ in task correctness over 282 non-abstaining selections; the confidence tie-break versus the feasible cost–ID control is -0.0035 $[-0.0110, 0.0000]$ in supportedness. These are evaluation-only contrasts, not human calibration or evidence of a uniform ranking benefit.

On those covered rows, the TAFER-minus-unconstrained contrast in supportedness was -0.0077 with interval $[-0.0528, 0.0356]$ for HotpotQA, $+0.0638$ $[+0.0288, +0.1070]$ for FEVER, $+0.0338$ $[-0.0068, +0.0811]$ for SciFact, and $+0.0523$ $[0.0000, +0.1071]$ for Mistral. The corresponding task-policy success contrasts were $+0.0026$ $[-0.0191, +0.0256]$, $+0.0255$ $[+0.0042, +0.0496]$, $+0.0405$ $[0.0000, +0.0878]$, and $+0.0116$ $[-0.0298, +0.0500]$. These are legacy evaluation-contract fields, not human support truth, and each interval is a pointwise 10,000-replicate source-cluster bootstrap. The feasible cost–ID control has zero or near-zero observed outcome contrast with the TAFER confidence tie-break on the joined rows. Thus the cleanest supported interpretation is FEVER-specific evidence for a feasibility-filtering contribution, with no uniform deterministic ranking advantage. Missing candidate truth is not imputed and these results do not replace the frozen tables.

The frozen FEVER task contract explicitly requests one of `SUPPORTS`, `REFUTES`, or `NOT ENOUGH INFO`, and states that the last is a task label rather than abstention. We verified the system contract and

Table 12: Separate V87 fresh-generation diagnostic. Each row uses unique instances and source-cluster percentile bootstrap intervals (2,000 draws; seed 20260919). Loss and components are operational-judge proxies for cited versus equal-count uncited removal. TextEM is uncited minus cited removal; for FEVER this is exact task-class accuracy, with its separate condition scores in Table 13. FEVER proxy-loss columns are not interpreted as abstention or semantic support loss.

Domain	n/sources	Loss proxy	Abst./assert.	TextEM Δ
HotpotQA	100/62	$+0.31$ $[+0.21, +0.41]$	$+0.20/+0.11$ abst./assert.	$+0.17$ $[+0.08, +0.27]$
2Wiki	100/60	$+0.20$ $[+0.11, +0.30]$	$+0.11/+0.09$ abst./assert.	$+0.10$ $[-0.01, +0.22]$
FEVER	80/67	Proxy loss not interpreted (NEI is a task label)		$+0.575$ $[+0.430, +0.718]$

Table 13: V87 FEVER task-class accuracy after a frozen-contract mapping audit. Each condition has 80 unique instances from 67 source clusters. Intervals are pointwise 95% source-cluster percentile bootstrap intervals (2,000 draws; seed 20260919). TextEM scores the exact class label independently of the abstain flag; AnsweredEM additionally requires `abstain=false`.

Condition	n/sources	TextEM [95% CI]	Coverage / AnsweredEM
Full	80/67	0.8375 $[0.7342, 0.9367]$	$1.00/0.8375$
Sham repeat	80/67	0.8375 $[0.7342, 0.9367]$	$1.00/0.8375$
Cited removed	80/67	0.2625 $[0.1447, 0.3827]$	$1.00/0.2625$
Equal-count uncited removed	80/67	0.8375 $[0.7349, 0.9375]$	$1.00/0.8375$
Cited only	80/67	0.8625 $[0.7701, 0.9500]$	$1.00/0.8625$
Citation IDs renamed	80/67	0.8375 $[0.7342, 0.9367]$	$1.00/0.8375$
No evidence	80/67	0.2000 $[0.0921, 0.3095]$	$1.00/0.2000$

structured gold-free user payload for all 560 planned FEVER requests, matched all 80 unique instance IDs and full-condition queries to the local evaluation labels, and checked generation/job payload hashes. Mapping the exact answer classes to the frozen labels makes task TextEM estimable; Table 13 reports every condition. All 560 decisions set `abstain=false`, so coverage is 1 and TextEM equals AnsweredEM; 58/80 cited-removal outputs are the task class `NOT ENOUGH INFO`, not abstentions. This repairs the prior task-score omission only: task correctness is not evidence-support validity, and the operational support-judge output remains a proxy. The historical response-mode figure is unchanged.

Fresh V87 generation contrasts. This is a separate, post-generation-gold-join diagnostic, not a replication of the historical condition-specific-generation analysis. HotpotQA and 2Wiki have complete paired samples of 100 unique instances with 62 and 60 source clusters, respectively. The reported source-cluster percentile intervals use 2,000 resamples and seed 20260919; they are pointwise. For objective task behavior, TextEM is scored independently of the abstention field, while AnsweredEM also requires a non-abstaining decision. For natural-language support, the judge output remains an operational proxy. The frozen scorer preserves 13 answer/abstention field conflicts and one failed generation rather than normalizing either. Table 12 reports the pre-specified cited-versus-equal-count-uncited contrast. The positive 2Wiki TextEM point estimate is uncertain because its interval includes zero. Fresh FEVER task EM is now reported in Table 13: the three exact response classes match the frozen contract, and raw `NEI` labels are mapped explicitly to the evaluation artifact's `NOT_ENOUGH_INFO` representation after an 80/80 ID, query, and label-lineage check. All FEVER outputs set `abstain=false`; its 58/80 `NOT ENOUGH INFO` cited-removal outputs are task-class predictions, not abstentions. This objective task score does not validate natural-language support, and the algebraic operational-proxy decomposition is not interpreted as semantic support loss.

Fresh V87 execution and missingness. After the initial 25-request stop (24 valid outputs) and one bounded diagnostic retry, the lane reached 1,959/1,960 valid generations and 1,420/1,421 valid operational-support responses. The failed generation was HotpotQA sham-repeat (99/100); the quarantined support response was HotpotQA citation-ID-renamed (99/100; 62 sources). Primary cited-versus-uncited support contrasts are complete for HotpotQA (100/100; 62 sources) and 2Wiki (100/100; 60 sources). Passed requests were not replayed. All results are pair-specific operational diagnostics, separate from historical comparisons and without human support validation.

Same-prompt sham negative control. We recomputed full–sham, full–cited-removal, and full–uncited-removal TextEM contrasts from the retained scored rows. Estimates below use the common four-condition cohort for each domain, so loss gaps share the same instances; losses are defined as $\mathrm{TextEM}_{\mathrm{full}} - \mathrm{TextEM}_{c}$. The cited-loss-minus-sham contrast is therefore $\mathrm{TextEM}_{\mathrm{sham}} - \mathrm{TextEM}_{\mathrm{cited}}$. The repeated full prompt is a sham control, not an independent seed or replication. Normalized answer strings changed in 8.1% of HotpotQA and 8.0% of 2Wiki full–sham pairs; string changes are not semantic errors.

For the same cohorts, cited-loss-minus-uncited-loss is $+.172$ $[+.082, +.267]$ in HotpotQA and $+.100$ $[-.010, +.217]$ in 2Wiki. Thus the task-correctness control supports a bounded HotpotQA contrast; the corresponding 2Wiki intervals include zero. The generic scored-row artifact has no FEVER TextEM/AnsweredEM values, so FEVER is not estimated in this table; its separately audited cited-versus-uncited task-class result is reported in Table 13. All intervals are pointwise, and no multiplicity-adjusted claim is made.

An exploratory source-equal reweighting of the same common cohorts gives a cited-loss-minus-sham-loss TextEM contrast of $+.169$ $[+.073, +.274]$ for HotpotQA and $+.092$ $[-.025, +.208]$ for 2Wiki. Equal-source leave-one-out point estimates range from $+.156$ to $+.189$ and from $+.076$ to $+.110$, respectively; these ranges are descriptive, not confidence intervals. The 2Wiki bootstrap interval still includes zero. This post-hoc sensitivity leaves the instance-weighted analysis primary and does not change the support-validity boundary.

An endpoint crosswalk on the same retained V87 scored rows found zero mismatches with the frozen AnsweredEM rule (TextEM correct and `abstain=false`) wherever TextEM was observed. On the common four-condition cohorts, cited-loss-minus-sham-loss point estimates were identical for TextEM and AnsweredEM: $+.172$ for HotpotQA (99 instances, 62 sources) and $+.110$ for 2Wiki (100 instances, 60 sources); source-cluster intervals were similar, with both 2Wiki intervals including zero. The unconditional abstention-rate increase was $+.343$ $[+.234, +.451]$ for HotpotQA and $+.130$ $[+.010, +.248]$ for 2Wiki. Separately, the structural audit counted 13 condition-level rows across the retained artifact with `abstain=true` and a nonempty answer string outside the frozen abstention-sentinel vocabulary (9 HotpotQA, 4 2Wiki). Within cited removal versus sham, these counts were 3 versus 0 for HotpotQA and 1 versus 0 for 2Wiki; the corresponding conflict-rate changes were $+.030$ $[.000, +.067]$ and $+.010$ $[.000, +.031]$, with both intervals including zero. These are string/flag conflicts, not semantic assessments, and no output was recoded. FEVER TextEM and AnsweredEM remain unavailable in this generic scored-row artifact; its separately mapped task-label evaluation is reported above. The crosswalk is an operational task-scoring diagnostic, not support-validity evidence.

Table 14: Secondary V87 sham-control TextEM reanalysis on common four-condition cohorts. Entries are signed paired differences with pointwise 95% source-cluster bootstrap intervals (10,000 draws; seed 20260924). Positive loss gaps indicate greater TextEM loss under cited removal. This is a bounded diagnostic, not a causal or support-validity result.

Domain	n/sources	Contrast	Estimate [95% CI]
HotpotQA	99/62	Full−sham	$+.010$ $[.000, +.032]$
HotpotQA	99/62	Full−cited	$+.182$ $[+.091, +.276]$
HotpotQA	99/62	Full−uncited	$+.010$ $[-.021, +.048]$
HotpotQA	99/62	Cited−sham loss	$+.172$ $[+.080, +.273]$
2Wiki	100/60	Full−sham	$-.010$ $[-.044, +.021]$
2Wiki	100/60	Full−cited	$+.100$ $[-.011, +.216]$
2Wiki	100/60	Full−uncited	$.000$ $[-.069, +.068]$
2Wiki	100/60	Cited−sham loss	$+.110$ $[-.010, +.232]$

Table 15: Controlled Stage3 comparison with a retrospectively reconciled local replay; the historical remote source hash is unavailable. Coverage is the answered fraction. FGR and WG-FGR are operational-checker quantities; "0 obs." denotes no observed checker-derived events under the tested protocol, not a population guarantee.

Method	FGR	WG-FGR	Coverage	Policy succ.	Tail mean90
Raw	.1207	.4250	.5814	.6240	9.5000
Self-correction	.0679	.2450	.6471	.7819	9.1709
TAFER $K = 4$	0 obs.	0 obs.	.5829	.8686	1.0000

Table 16: Historical audit. The primary all-category A/B/C analysis is shown first (150 pairs). A/B only is a post-hoc secondary complete-case sensitivity excluding 70 pairs where either rater chose C (no cluster interval estimated). Certificate flags test structure, not semantic support; this audit does not cover fresh outputs.

Regime	Paired n	Agreement	κ	Known cert.
Controlled Qwen	25	0.520	0.068	0
Qwen HotpotQA	35	0.714	0.572	30
Qwen FEVER	30	0.467	0.140	22
Qwen SciFact	25	0.600	0.424	0
Mistral HotpotQA	20	0.550	0.297	0
Qwen 2Wiki	15	0.400	0.049	8
All (primary)	150	0.560	0.333	60
A/B only (secondary; post-hoc)	80	0.788	0.558	–

The rule-task response-type audit is exploratory. In explicit single-proof anchor-removal worlds, 117 of 120 outputs assert the distractor subject's badge and three abstain. In the implicit counterpart, 42 assert that distractor badge and 78 abstain. The rule is explicit in only one instruction condition, so an implicit-condition score alone cannot establish failure to follow a supplied rule. Programmatic answerability and matched deletion counts were rechecked for every condition. Earlier rule-task versions, including the quarantined implementation defect, are not combined with this task.

M HISTORICAL HUMAN AUDIT AND UNRESOLVED CALIBRATION

The author supplied 180 labels. The current historical audit uses a 150-choice email return joined to the author subset through role-isolated canonical maps. A preceding 150-choice return, completed after the annotator had seen disagreement cases, yielded 87/150 agreement (0.580; three-category kappa 0.369). A later audit verified that its public pair set matches the private canonical maps; this post-exposure return is not a blind independent replication. We preserve but do not pool it; the later email return yields 84/150 agreement (0.560; kappa 0.333). The author reports that the annotator read without AI assistance or access to the author's labels, but the retained email contains no direct annotator process attestation. Because the preceding return followed exposure to disagreement cases, the records do not establish whether the later map-verified return was blind to that exposure or the annotator's earlier judgments. We therefore report 84/150 as historical inter-rater agreement, not a blind replication or fresh-output calibration; identity, blinding and exposure remain author-reported, not independently verified. The author's own prior exposure to AI-assisted materials also limits interpretation.

An exposure-only exact-text audit compared the historical and prepared V87 packet projections of question, visible evidence and answer, without labels, IDs or private maps. In the independent packet, prior question text recurred in 20/50 canonical question groups (60/150 condition rows); 3/150 rows matched on both question and visible evidence, and 0/150 matched on question, evidence and answer. In the author packet, question text recurred in 29/60 groups (87/180 rows); 3/180 rows matched on question and evidence, and 0/180 matched on all three fields. The V87 packets were blank at the audited snapshot. These counts describe exact overlap detectable in those packet fields only; they do not establish annotator identity, independence, what other material was seen, or absence of semantic exposure. A later return from the same annotator must not be described as wholly unexposed.

Role-isolated public IDs map privately to 150 canonical pairs. The exact structural Excel projection separates question and evidence without changing output or labels. Six FEVER rows have empty evidence; they remain in the audit but cannot validate complete natural-language support contexts. No automated label inference or adjudication was performed.

A denotes supported, B unsupported, and C uncertain. Main agreement and kappa retain all three categories. Estimates describe the audit sample, not an inclusion-weighted paper population. There are 143 dataset/source clusters; 7 HotpotQA sources occur in both architectures. These rows are resampled jointly in 10,000 dataset-stratified cluster draws (seed 17), fixing observed dataset allocation. Intervals are pointwise, not simultaneous.

For the current map-verified return, 84/150 labels agree, with 66 disagreements. Raw agreement is 0.5600, with cluster interval [0.4826, 0.6378]; three-category kappa is 0.3327, with interval [0.2177, 0.4453]. Category-specific positive and negative agreement are 0.6349 and 0.5542. Either rater is uncertain on 70/150 pairs. This return contains 55 A, 46 B and 49 C responses.

For AR3-010 and AR3-024, the author later clarified the rating object: support applies only to substantive assertions; a pure abstention or a question echo without a substantive claim is not applicable, and abstention appropriateness is outside this support label. This item-specific clarification does not recode the historical responses or adjudicate either case.

Excluding only these two author-designated non-applicable items, while retaining all recorded votes, gives 84/148 agreements (0.5676; three-category kappa 0.3429), with 68/148 pairs containing an uncertain response. The dataset-stratified source-cluster 95% intervals are [0.4893, 0.6455] for agreement and [0.2276, 0.4566] for kappa (10,000 draws, seed 17). This narrow, post-hoc scope sensitivity does not replace the 84/150 all-item primary result or constitute adjudication. The broader structural applicability sensitivity below is a separate estimand.

The author selected a global support-only rule: a response is not applicable when its output contains no substantive answer, including explicit structured abstentions, exact literal no-answer outputs, and malformed outputs with exact no-answer/abstain fields and no claim field. Classification uses output structure, not either rater's label; no semantic alias mapping is used. This post-hoc rule excludes 58/150 rows (25 structured, 24 literal, and 9 malformed-explicit), leaving 59/92 agreements (0.6413; descriptive Wilson interval [0.5395, 0.7318]) and three-category kappa 0.3912. Across 88 dataset/source clusters, 10,000 dataset-stratified source-cluster draws (seed 17) give pointwise intervals [0.5444, 0.7340] for agreement and [0.2444, 0.5306] for kappa. No historical vote changes. This broader structural applicability analysis is distinct from the two-item scope sensitivity above. Neither post-hoc sensitivity replaces the current 84/150 descriptive agreement record, resolves the independent rater's intent, scores abstention appropriateness, or validates fresh V87 outputs.

Excluding uncertain pairs leaves 80 pairs, with 63 agreements (0.7875) and kappa 0.5584. This outcome-conditioned secondary analysis must not replace the full-sample gate or be used to discard disagreement. Prespecified internal strong (agreement 0.80, kappa 0.60) and bounded (0.70, 0.40) gates both fail on all 150 pairs. These thresholds are not ICLR requirements and do not themselves certify support truth.

The linked field was exported as `checker_support`, but all 60 known values originate from `certificate_pass`: a structural schema/citation/abstention check, not a Qwen semantic judgment. The 90 absent values remain unknown. We use the independent return as a single-rater reference, not adjudicated truth; the quantities below are reference-relative disagreements. C responses are reported explicitly. For certificate-passing rows, disagreement is B; for certificate-failing rows, disagreement is A. Definite-only rates condition on A/B. The uncertainty range over all linked rows is $[e/n, (e+u)/n]$, with e definite disagreements and u uncertain labels; it is an identification range, not a confidence interval.

For certificate-passing outputs, there are 36 linked rows, 8 uncertain judgments and 11/28 definite disagreements (0.3929). The uncertainty identification range is [0.3056, 0.5278]. The descriptive Wilson interval on definite judgments is [0.2357, 0.5759], conditional on this observed subset, not the full paper population.

For certificate-failing outputs, there are 24 linked rows, 4 uncertain judgments and 15/20 definite disagreements (0.7500). The uncertainty identification range is [0.6250, 0.7917]. The descriptive

Table 17: Structural certificate flags versus the independent return. Each cell is definite disagreements / definite judgments (uncertain judgments). A dash means no linked flag. Missingness prevents population calibration.

Regime	Known / missing cert.	Passing cert.	Failing cert.
Controlled Qwen	0/25	–	–
Qwen HotpotQA	30/5	6/15 (4)	8/10 (1)
Qwen FEVER	22/8	2/10 (3)	6/7 (2)
Qwen SciFact	0/25	–	–
Mistral HotpotQA	0/20	–	–
Qwen 2Wiki	8/7	3/3 (1)	1/3 (1)

Wilson interval on definite judgments is [0.5313, 0.8881], conditional on this observed subset, not the full paper population.

The retained schema/citation checker reproduces all 60 stored flags; it does not assess entailment. Twenty-two rows lack `schema_valid`; reparsing the same raw strings would change 20 flags, all raw candidates. The original branch did not reparse; its flags remain. The 90 absent flags in the table remain unknown. These are structural-flag disagreements, not semantic-judge calibration. The Qwen/DeepSeek comparison uses a different subset (Appendix N). Packet hashes and retained-code replay anchor lineage; the original runtime revision is unverified. Fresh intervention effects remain uncalibrated. A blank disagreement packet is retained; original labels remain unadjudicated.

M.1 Second-pass review of the original disagreements

The same independent annotator re-reviewed all 66 disagreements in the later canonical-map-paired 150-choice return. Both earlier label sets were hidden, items were reordered under a fixed hash, and new public IDs were used. The original question, evidence and output strings were preserved exactly, but long text was displayed in continuation rows. This is a selected within-rater follow-up, not a new independent sample, a randomized layout comparison or human adjudication. Prior exposure may still affect later judgments.

The return contains 44 A, 22 B and 0 C responses. The annotator changed 33/66 judgments and retained 33/66. Of the re-reviewed items, 25 now match the unchanged author response and 41 still differ. Only 4 of the 17 original definite A/B disagreements now match the author. These counts do not establish which judgment is correct. No rationale or joint adjudication was returned.

Table 18: Second-pass follow-up, conditional on disagreement in the later canonical-map-paired return. Matching the author is not an adjudicated correctness criterion. Its 84/150 agreement and kappa are unchanged.

Regime	Reviewed	Changed	Now matches author	Still differs
Controlled Qwen	12	7	7	5
Qwen HotpotQA	10	2	1	9
Qwen FEVER	16	11	7	9
Qwen SciFact	10	6	4	6
Mistral HotpotQA	9	2	1	8
Qwen 2Wiki	9	5	5	4

All 28 initially uncertain judgments in this selected subset became definite: 26 changed to A and two to B. Of the 15 initial A judgments, 14 stayed A and one became B; of the 23 initial B judgments, 19 stayed B and four became A. No previously concordant item was re-reviewed. The combined count below is therefore explicitly a sequential, selected re-review summary, not full-sample retest reliability.

Four re-reviewed items had no displayed evidence; all received A on this pass. The workbook defined A as visible-evidence support for a substantive output claim; it did not separately ask whether abstention was appropriate. Two outputs are the literal string INSUFFICIENT_EVIDENCE and two are JSON objects declaring abstain=true. Under the author's global rule, all four are outside

support scoring, as are structurally equivalent non-answer abstentions elsewhere in the historical audit. This scope decision does not recode the recorded human labels or assess whether abstention was appropriate.

The author subsequently re-reviewed the 41 items still differing from the independent annotator's second pass, using new public IDs and the original displayed text, without either prior label in the packet. The author supplied 22 A and 19 B responses: 40 changed from the author's first pass, 40 matched the independent second pass, and one still differed. This was a selected third look by a non-independent rater, not a consensus procedure or a second random sample. For AR3-010 and AR3-024, the author later clarified that the support label applies only to substantive assertions; a pure abstention or non-substantive question echo is not applicable, and abstention appropriateness is outside the label. This does not recode their historical responses or adjudicate the cases. The original 150-pair agreement, kappa, uncertain labels and unadjudicated status are unchanged.

Across the 66 initially discordant cases, 25 matched on the independent annotator's second pass; the remaining 41 were re-reviewed by the author. The initial author recheck matched 40/41; the later AR3-012 response matches the independent second-pass response, giving 66/66 selected sequential alignment in the current working record. Under the global support-only rule, 33 of these cases are applicable: two aligned on the independent second pass and 31/31 author-rechecked applicable cases align after the later response. The other 33 are outside support scoring. Together with 84 initially concordant pairs that were not re-reviewed, this remains selected sequential convergence, not a replacement reliability statistic, formal adjudication, or validated support truth.

The current canonical-map-paired 84/150 agreement, three-category kappa 0.3327, uncertainty counts and calibration denominators remain the descriptive audit record. Selective re-review cannot replace those estimates or remove Qwen-family judge dependence. Natural-language support and historical-to-fresh-output transfer remain unvalidated.

N DIFFERENT-FAMILY JUDGE SENSITIVITY

Question and frozen design. Does a bounded result depend on the operational evaluator? We compare Qwen-Turbo with DeepSeek-Flash (the provider's DeepSeek-V4.1-Flash alias), not with a new answer generator. We freeze the sample, binary evidence-support rubric and endpoints before querying. The rubric and visible question/evidence/answer messages are identical across evaluators. No human judgment, task gold, generator identity or previous checker decision is supplied. Non-thinking mode and temperature zero are used once per distinct payload; identical payloads are reused, not counted as independent replications. Both models can share errors.

Natural-output branch. Frozen metadata hashes select 30 unique instances in each of HotpotQA, FEVER and 2Wiki, without reading support outcomes and without replacing missing rows. The same existing full, cited-removed and uncited-removed generations are used. Historical Qwen judgments are reused; DeepSeek receives exactly the same messages. Generation abstention is a separate observed decision, not an inferred support label. For condition c, define $U_i^c = 1$ only for a non-abstaining answer rejected by the operational judge; abstentions contribute zero to this assertion frequency. The primary contrast is $\Delta_U = \overline{U^{cited} - U^{uncited}}$ over complete paired instances. This is not false-green risk because objective correctness is not part of this endpoint. All 30 selected pairs in each domain are complete. Intervals resample source clusters 10,000 times (seed 17), rather than treating the three outputs per instance as independent units.

Domain	n	Sources	Qwen Δ_U [95% CI]	DeepSeek Δ_U [95% CI]
HotpotQA	30	27	-0.033 $[-0.207, +0.138]$	$+0.133$ $[-0.036, +0.321]$
FEVER	30	28	$+0.633$ $[+0.452, +0.821]$	$+0.567$ $[+0.379, +0.759]$
2Wiki	30	25	$+0.100$ $[-0.040, +0.267]$	-0.033 $[-0.148, +0.071]$

Table 19: Frozen-output cross-judge sensitivity, not human support truth. Uncertain HotpotQA and 2Wiki point estimates reverse sign; FEVER retains its direction. Fresh FEVER emits task-class strings including NEI with an explicit non-abstention flag, so this contrast must not be pooled with the historical abstention interface.

The paired DeepSeek-minus-Qwen contrast is $+0.167$ $[+0.000, +0.345]$ for HotpotQA, -0.067 $[-0.258, +0.103]$ for FEVER, -0.133 $[-0.355, +0.071]$ for 2Wiki. The same underlying answers therefore do not give uniformly stable operational conclusions. A secondary directional full-supported-to-unsupported loss is evaluated only on all-asserting complete triplets (20/30/13 instances and 19/28/12 source clusters). Its Qwen versus DeepSeek point estimates are 0.150 versus 0.300 on HotpotQA, 0.567 versus 0.567 on FEVER, and 0.231 versus -0.077 on 2Wiki. This selected secondary population does not replace the primary comparison.

Historical independent-reference branch. Both models also rate the original 150 historical audit objects, excluding six rows whose displayed evidence is empty. All original labels and uncertainty remain unchanged; the disagreement-selected second review is not substituted. The primary reference comparison requires an explicitly non-abstaining output and a definite original independent rating. Exact JSON parsing identifies 75 non-abstaining outputs; 73 have definite A/B ratings. Abstention-form outputs and unresolved formats are retained separately, not semantically relabeled. No training, threshold fitting, adjudication or population weighting is performed. On these 73 items (72 dataset/source clusters), reference agreement is 53/73 (0.726, interval [0.625,0.822]) for Qwen and 61/73 (0.836, [0.753,0.917]) for DeepSeek. The paired difference is 0.110 [0.014,0.205]. Both judges disagree with the reference on nine common items. This comparison is with one original independent annotator, not established support truth. The judges themselves disagree on 14 of these 73 decisions. In the broader 100-item comparison, they disagree on 23 decisions and both differ from the reference on 10 common items. The family sensitivity therefore changes some item-level calls but supplies neither a validated criterion nor evidence of independent model errors.

Including every definite rating with displayed evidence gives 100 items and a smaller agreement difference: Qwen 0.770 versus DeepSeek 0.800, paired difference 0.030 [-0.049,0.113]. This secondary scope mixes rating objects, including abstention. The observed model comparison is therefore sensitive to the declared audit scope; the stronger subset result cannot be generalized to all

Table 20: Retrospective joint-cohort sensitivity on the fixed cross-judge sample. Entries are paired differences with pointwise 95% source-cluster percentile intervals (10,000 draws; seed 17). TextEM/AnsweredEM are uncited minus cited; abstention and operational unsupported-assertion proxy frequencies are cited minus uncited. The final column is the DeepSeek-minus-Qwen difference in those proxy contrasts. FEVER's zero abstention contrast is constant in this sample; its degenerate interval is descriptive, not evidence of equivalence.

Domain (n/src.)	Contrast	Estimand	Estimate [95% CI]
HotpotQA (30/27)	TextEM	uncited − cited	$+.133$ $[+.031, +.276]$
HotpotQA (30/27)	AnsweredEM	uncited − cited	$+.133$ $[+.031, +.276]$
HotpotQA (30/27)	Abstention	cited − uncited	$+.267$ $[+.103, +.448]$
HotpotQA (30/27)	Qwen proxy	cited − uncited	$-.033$ $[-.207, +.138]$
HotpotQA (30/27)	DeepSeek proxy	cited − uncited	$+.133$ $[-.036, +.321]$
HotpotQA (30/27)	Judge gap	DeepSeek − Qwen	$+.167$ $[.000, +.345]$
FEVER (30/28)	TextEM	uncited − cited	$+.633$ $[+.452, +.821]$
FEVER (30/28)	AnsweredEM	uncited − cited	$+.633$ $[+.452, +.821]$
FEVER (30/28)	Abstention	cited − uncited	$.000$ $[.000, .000]$
FEVER (30/28)	Qwen proxy	cited − uncited	$+.633$ $[+.452, +.821]$
FEVER (30/28)	DeepSeek proxy	cited − uncited	$+.567$ $[+.379, +.759]$
FEVER (30/28)	Judge gap	DeepSeek − Qwen	$-.067$ $[-.258, +.103]$
2Wiki (30/25)	TextEM	uncited − cited	$+.133$ $[.000, +.300]$
2Wiki (30/25)	AnsweredEM	uncited − cited	$+.133$ $[.000, +.300]$
2Wiki (30/25)	Abstention	cited − uncited	$+.033$ $[.000, +.100]$
2Wiki (30/25)	Qwen proxy	cited − uncited	$+.100$ $[-.040, +.267]$
2Wiki (30/25)	DeepSeek proxy	cited − uncited	$-.033$ $[-.148, +.071]$
2Wiki (30/25)	Judge gap	DeepSeek − Qwen	$-.133$ $[-.355, +.071]$

outputs. Original author–independent agreement remains 0.5600 and is not replaced by model–human agreement.

Interpretation and provenance. This analysis reduces dependence on an unexamined single evaluator, but does not identify or eliminate same-family preference. DeepSeek accepts more outputs in the 73-item audit subset (60/73 versus 54/73); the comparison is not evidence that Qwen simply favors its own answers. Historical Qwen and new DeepSeek evaluations were obtained at different times, and provider model names are aliases rather than independently verified weight snapshots. Natural-output proxy conclusions remain bounded; the historical human audit does not validate these fresh answers. All null, reversed and discordant results are retained. No generator, original candidate pool, historical label, primary paper table or registered abstract was changed.

Joint task and judge endpoints on the same fixed cohort. We additionally joined the same 30 hash-selected instances per domain across the full, cited-removed and uncited-removed generations, task scores, abstention decisions, and both operational judges. FEVER task-class scores were recomputed from the frozen post-generation gold join using the separately audited label mapping; the official target remains fixed across evidence conditions. In particular, NOT ENOUGH INFO is a task class, not an abstention, and a change in class match is not by itself a support-validity judgment. TextEM and AnsweredEM are shown separately even where they coincide. Each endpoint uses all 30 instances and source-cluster bootstrap weights shared across endpoints within domain; source IDs are never inferred from instance IDs.

In this small fixed subset, the FEVER task-class and both judges' proxy contrasts point in the same direction. The HotpotQA and 2Wiki proxy estimates have judge-dependent signs and intervals spanning zero; the paired judge contrast also does not establish a general difference between evaluators. These are retrospective diagnostics, not confirmatory replications, human calibration, semantic support labels, or evidence that judge errors are independent. They supplement, but do not replace, the endpoint-specific analyses on their original complete-case denominators.

O ABSTENTION, MISSING EVALUATION FIELDS, AND POLICY YIELD

Historical information boundary. This audit used only the support fields attached to previously selected outputs. It did not yet execute the original citation-set scorer on every pool candidate. Appendix Q subsequently recovers unique instance-level evaluation metadata and completes that legacy relation for all actual candidates. The bounds below preserve the earlier information set; they are not a claim that those citation-set fields remain irrecoverable, nor bounds on human support truth.

Why the denominators differ. The exact-selector replay joins evaluation fields after observable-only selection. A policy abstention has no selected candidate to join; this is not the same as a selected answer with a missing support field. In the four frozen reference pools, TAFER returns 390/468, 235/481, 148/202 and 172/222 answers. All these selected answers have legacy supportedness fields. The remaining 78/246/54/50 instances are abstentions. The same-key unconstrained policy instead has 0/2/0/1 selected answers with missing fields; these remain unknown.

An all-instance endpoint, not an abstention label. Let p_{ij} be the selector's probability of returning candidate j for instance i, and let v_{ij} denote the stored evaluation support field. An abstaining policy has $\sum_j p_{ij} = 0$, while an answering policy has sum one. Define returned-supported-answer yield as $Y(p) = N^{-1} \sum_{i,j} p_{ij} v_{ij}$. An abstention returns no supported answer and contributes zero to this endpoint; we do not label the abstention unsupported or incorrect. This endpoint differs from task-policy success, conditional answer quality and false-green loss. Legacy v is not human support truth, and this audit does not establish FEVER EM.

For two policies p, q, the contrast has coefficients $w_{ij} = p_{ij} - q_{ij}$. Known fields contribute $w_{ij} v_{ij}$; a missing binary field contributes an interval $[\min(0, w_{ij}), \max(0, w_{ij})]$. Summing yields sharp algebraic bounds over the missing fields. The same missing candidate is shared across policies, so identical coefficients cancel before bounding. These finite-cohort ranges are not confidence intervals. No missing labels are filled. Uniform feasible selection uses its exact probability $1/|\mathcal{F}_i|$ on the same feasible pool, not repeated random draws counted as additional data.

Regime	N	Sources	Δ coverage	Filter ΔY range	Ranking ΔY range
Qwen HotpotQA	468	78	-0.0342	$[-0.0107, -0.0107]$	$[-0.1190, +0.1020]$
Qwen FEVER	481	105	-0.2536	$[+0.0270, +0.0312]$	$[-0.0078, +0.1219]$
Qwen SciFact	202	202	-0.0545	$[+0.0198, +0.0198]$	$[-0.0876, +0.0912]$
Mistral HotpotQA	222	37	-0.0586	$[+0.0360, +0.0405]$	$[-0.1417, +0.0383]$

Table 21: Retrospective all-instance CPU sensitivity on unchanged named reference pools. Filter contrast: TAFER minus the unconstrained selector with the same cost/confidence/ID key. Ranking contrast: TAFER minus the analytic expectation of uniform feasible selection. The latter has identical answer coverage. Ranges propagate absent legacy candidate fields and are not sampling intervals.

The filter yield difference is point-identified for HotpotQA and SciFact. Their 10,000-draw source-cluster intervals are $[-0.0491, 0.0256]$ and $[-0.0149, 0.0545]$, respectively (seed 20260919). FEVER and Mistral retain missing-field bounds rather than an imputed effect and confidence interval. The observed positive finite-cohort FEVER range does not imply a population improvement. It must be considered alongside the 25.36-point coverage decrease.

TAFER versus uniform-feasible yield remains unidentified in 326/190/116/117 instances across these four pools because not every feasible candidate was evaluated. All four full-cohort ranges include zero. This expected-policy analysis neither overwrites the frozen sampled-random baseline nor establishes a ranking advantage. Feasible cost/ID comparisons likewise retain their 17/16/13/17 missing selected-candidate fields. The original complete-case contrasts reconcile exactly and are not replaced. Mistral's adverse TAFER-minus-raw FGR point estimate remains unchanged: its endpoint and comparator differ from the supported-answer yield comparison here.

No unique 2Wiki named-reference pool was substituted from another lane; controlled Qwen remains aggregate-only for candidate accounting. These are additional CPU analyses of existing observations, not new samples, model runs, or human validation.

P HISTORICAL NEGATIVE CONTROLS AND JOINT-INTERVAL SENSITIVITY

This retrospective CPU analysis adds no requests or observations and does not replace the original pointwise intervals. Before recomputation we fixed three contrasts (cited minus uncited, cited minus sham, renamed minus sham), two supported-answer-loss components, and three domains: 18 endpoints. This is not a new preregistration or a correction for every historical test in the paper. Every endpoint uses its own complete full/left/right triple; full and sham prompts match byte-for-byte on all 323 planned instances. HotpotQA has 129 planned triples and 128 complete triples (69 sources); FEVER has 82/82 (69), and 2Wiki 112/112 (63). No missing generation or required proxy is imputed. Initially unsupported or abstaining full outputs contribute zero supported-answer loss, as in the main definition; this does not relabel an abstention unsupported.

We draw source clusters with replacement within each domain, sharing the same cluster weights across its endpoints. Each draw's estimate is the sum of weighted instance outcomes divided by its corresponding weighted complete-case count. For 10,000 draws (seed 20260923), let $\hat{\theta}_j$ be an endpoint mean, $\hat{\theta}^*_{bj}$ its bootstrap value, and s_j its bootstrap standard deviation. The reported joint band uses the empirical 95th percentile q of $\max_j |(\hat{\theta}^*_{bj} - \hat{\theta}_j)/s_j|$ and endpoints $[\hat{\theta}_j - qs_j, \hat{\theta}_j + qs_j]$, intersected with $[-1, 1]$. These are approximate centered max-standardized-deviation bands, not an exact coverage guarantee. The estimated scale is fixed across draws, not a nested studentized bootstrap. Source independence and domain-stratified sampling remain assumptions. Seventeen endpoints have nonzero bootstrap variation. The remaining HotpotQA renamed-minus-sham abstention component is identically zero in the observed cohort and is marked non-estimable rather than assigned an equality claim.

Domain	Contrast	ΔB	Joint band	ΔU	Joint band
HotpotQA	Cited–uncited	$+0.172$	$[+0.053, +0.291]$	$+0.148$	$[+0.022, +0.274]$
HotpotQA	Cited–sham	$+0.195$	$[+0.074, +0.316]$	$+0.180$	$[+0.063, +0.297]$
HotpotQA	Renamed–sham	$+0.000$	NE	$+0.023$	$[-0.017, +0.064]$
FEVER	Cited–uncited	$+0.561$	$[+0.379, +0.743]$	$+0.012$	$[-0.069, +0.094]$
FEVER	Cited–sham	$+0.573$	$[+0.389, +0.757]$	$+0.049$	$[-0.021, +0.119]$
FEVER	Renamed–sham	$+0.049$	$[-0.038, +0.136]$	$+0.037$	$[-0.024, +0.097]$
2Wiki	Cited–uncited	$+0.152$	$[+0.023, +0.281]$	$+0.054$	$[-0.045, +0.152]$
2Wiki	Cited–sham	$+0.152$	$[+0.034, +0.270]$	$+0.098$	$[-0.006, +0.202]$
2Wiki	Renamed–sham	-0.009	$[-0.036, +0.018]$	$+0.054$	$[-0.009, +0.116]$

Table 22: Historical operational-proxy loss components: B denotes loss through abstention and U loss through continued assertion. Joint bands cover the 17 estimable endpoints of this declared sensitivity family, approximately under the bootstrap assumptions; NE denotes no observed bootstrap variation. No human truth, FEVER EM, causal effect or population safety claim follows.

HotpotQA's two cited-removal components and all three domains' abstention components retain positive joint bands against both uncited and sham controls. FEVER and 2Wiki continued-assertion contrasts with sham have positive pointwise percentile intervals but joint bands containing zero. The same qualification applies to 2Wiki's renamed-minus-sham continued-assertion component. These weakened comparisons are retained, not selected away. Every estimable renaming-versus-sham joint band contains zero; that is not evidence of equivalence or a validated ID-invariance guarantee.

Leave-one-source-out and equal-source-weight summaries are retained in local audit records and are not included in the current submission files. Removing one source does not reverse the cited-minus-uncited point-estimate signs for the six historical components, although FEVER's small continued-assertion component can become zero. This is an influence check, not a confidence statement or robustness to unknown support-label errors. The historical point estimates, instance counts and source counts reconcile exactly with the unchanged main figure. Different-family evaluator sensitivity uses fresh outputs and must not be pooled with these historical records.

Q COMPLETE-CANDIDATE REPLAY OF THE STORED CITATION RELATION

The original selected-output tables evaluate policies, not every candidate in each pool. We recovered a unique stored gold-citation-ID set G_i for every instance and applied the original scoring function to each actual candidate's citation set Q_{ij}:

$$R_{ij}^{\mathrm{ID}} = \mathbb{1}[Q_{ij} \neq \varnothing \ \wedge \ Q_{ij} \subseteq G_i].$$

This is citation-set admissibility, not semantic entailment or completeness: even a wrong answer or an incomplete proof can pass. A recorded gold-support-ID set is not a complete natural-language logic contract. This CPU calculation does not activate Track B, change human labels, or resolve support validity.

Before joining G_i, each policy chooses using the unchanged observable cost/confidence/ID and feasibility fields. The full-cohort endpoint is $Y_{\mathrm{ID}}(\pi) = N^{-1} \sum_{i,j} p_{ij}^{\pi} R_{ij}^{\mathrm{ID}}$; abstention has no selected answer and contributes zero yield, not an incorrect or unsupported-abstention label. Uniform feasible uses exact selection probabilities over the same feasible pool. Its answer coverage equals TAFER's; Monte Carlo repetitions are not added observations.

Compatibility before recovery. All 12,357 original scored support fields are exactly reproduced by the retained function. Every candidate-linked evaluation has the same source and citation list as the frozen candidate. Conflicting metadata, missing explicit lists, payload drift or a nonreproduced score would block the affected calculation; the function's fallback to checker/certificate fields is not used. Metadata uniqueness is checked across every old policy row for an instance. This verifies compatibility with the recorded relation, not independent semantic accuracy. Across the four pools, all 8,238 candidates are now scored, including 2,615 candidates absent from the earlier selected-output lookup. The four candidate counts are 2,808/2,886/1,212/1,332 in the table's order. No candidate is generated, duplicated or inferred from a summary. Table 23 reports 10,000-draw pointwise source-cluster intervals (seed 20260919), not simultaneous or confirmatory guarantees.

Regime	N/sources	Filter ΔY_{ID}	Rank ΔY_{ID}
Qwen HotpotQA	468/78	-0.0107 $[-0.0491, +0.0256]$	-0.0144 $[-0.0381, +0.0079]$
Qwen FEVER	481/105	$+0.0312$ $[+0.0139, +0.0517]$	$+0.0005$ $[+0.0000, +0.0016]$
Qwen SciFact	202/202	$+0.0198$ $[-0.0149, +0.0545]$	$+0.0090$ $[-0.0092, +0.0280]$
Mistral HotpotQA	222/37	$+0.0405$ $[+0.0000, +0.0856]$	-0.0098 $[-0.0325, +0.0123]$

Table 23: All-instance citation-relation yield on the same unchanged pools. Filter: TAFER minus the unconstrained selector with identical cost/confidence/ID key. Rank: TAFER minus expected uniform-feasible selection. These are not human support, task-correctness or false-green-risk contrasts.

As a retrospective multiplicity sensitivity, an approximate stratified source-cluster max-$|t|$ bootstrap over the 12 domain-by-policy contrasts leaves only the FEVER filter-versus-unconstrained contrast excluding zero (simultaneous interval $[+0.0022, +0.0602]$). None of the other 10 estimable contrasts, including every TAFER-versus-ranking-control contrast, excludes zero. The FEVER TAFER-versus-feasible-cost-key contrast is exactly zero for every instance in the observed cohort and is therefore reported as a finite-cohort identity, not a population interval. The separate TAFER-versus-uniform-feasible estimate is $+0.00052$, with a simultaneous interval that includes zero. This retrospective sensitivity does not make the analysis confirmatory and concerns only the stored citation-set relation.

A same-cohort source-influence check finds no sign reversal for any nonzero contrast after deleting one source at a time. For FEVER filtering, the instance-weighted estimate is $+0.0312$, the equal-source estimate is $+0.0273$, and all 105 leave-one-source estimates lie between $+0.0252$ and $+0.0316$. These are post-hoc finite-cohort sensitivities, not confidence intervals or new observations.

Coverage and ranking qualification. TAFER coverage is 0.8333/0.4886/0.7327/0.7748 in the displayed order; filtering reduces coverage by 0.0342/0.2536/0.0545/0.0586 relative to the same-key unconstrained selector. The full-cohort FEVER yield contrast is now point identified at the upper end of its earlier missing-field range, not an improved experimental result. TAFER versus feasible cost/ID yield differs by -0.00427, 0, -0.00495 and -0.00450; all pointwise intervals contain zero. The FEVER pair is equal on this observed relation; its degenerate bootstrap interval does not prove

population equivalence. The $+0.00052$ uniform-feasible contrast is a mean expected-yield difference, not a binary event rate: on its sole contributing FEVER instance, TAFER yields 1.00 and the uniform mixture yields 0.75 across four feasible candidates, giving $0.25/481 = 0.00052$. This single-instance, single-source contribution does not establish a general ranking gain.

TAFER's mean returned recorded cost is lower than uniform feasible by 0.1323/0.0744/0.1069/0.1182 per instance. This follows its recorded-cost objective under type-assigned costs; it is not a measured edit-distance or utility gain. Mistral's original adverse TAFER-minus-raw FGR point estimate remains unchanged, with a different metric and comparator. Controlled Qwen has no recovered candidate-level pool; no 2Wiki pool from a different lineage is substituted. The earlier missing-field audit is retained for traceability, but missing selected-output support fields no longer prevent this particular citation-set comparison. Natural-language support remains uncalibrated by it.

2Wiki cross-lane lineage. The V82 full behavioral cohort contains 112 instances from 63 sources. The separately retained `qwen_twowiki400` candidate pool contains 2,400 candidates from 400 instances and 80 sources; its receipt identifies `qwen2.5-7b-instruct`, whereas the behavior protocol identifies `qwen-turbo`. Among the 112 shared instance/source pairs, no prompt hash, generation-job ID, or candidate ID matches. There are 27 normalized answer/citation overlaps, but only five are unique to one candidate row. Five additional pre-existing Qwen-Turbo diagnostic pools also have zero matching prompt hashes, generation-job IDs, or candidate IDs; their normalized answer/citation matches range from 26 to 33 rows, with zero to three unique matches. These content overlaps are descriptive only. No behavioral outcome is joined to a candidate-pool record, and no shared generation lineage is claimed.

R RETROSPECTIVE CANDIDATE-LEVEL TASK-UTILITY REAUDIT

This post-hoc CPU analysis replays five selectors on each frozen, named candidate pool. Candidate choices were fixed from observable pool fields before joining post-generation gold. We applied the existing task scorer to every candidate and checked all candidate-level score records that were present. Some pool candidates had no stored candidate-level score record; their outcomes were recomputed from the same instance-level gold and frozen scorer rather than imputed from a selector summary. The verified stored-score coverage and recomputed remainder are recorded in the accompanying audit receipt. This is a retrospective sensitivity analysis, not a new sample or a confirmatory test.

The primary contrast isolates feasibility filtering by holding the cost/confidence/ID tie-break fixed: deterministic TAFER minus the same-key non-abstaining selector without feasibility filtering. The second contrast isolates the confidence tie-break within the same feasible set by comparing TAFER with a cost/ID-only selector. A third contrast compares TAFER with the analytic expectation of uniform selection over that feasible set. Task-policy success uses the stored expected-action contract; returned answer coverage is reported separately. Support yield and false-green remain gold citation-set proxies, not semantic or human support truth.

The filtering contrast is positive in these three finite pools, although the HotpotQA interval reaches zero. The confidence tie-break point estimates are zero in HotpotQA and Mistral and slightly negative in SciFact; these intervals do not establish a nonzero tie-break effect. For TAFER minus analytic uniform-feasible task-policy success, the point estimate is negative in HotpotQA and Mistral but positive in SciFact; all three intervals include zero. This audit therefore does not establish an incremental ranking advantage, equivalence, or a general disadvantage. It supports only a bounded feasibility-filtering interpretation. The low Mistral task-policy success and the separate adverse Mistral gold citation-ID-proxy FGR point estimate remain visible; this endpoint does not reverse that negative result.

A source-equal reweighting check on the same frozen pools gives the same point estimates because each source contributes six instances in HotpotQA and Mistral, and one in SciFact. TAFER minus analytic uniform-feasible task-policy success is -0.0087 $[-0.0225, 0.0067]$ for HotpotQA, $+0.0046$ $[-0.0149, 0.0257]$ for SciFact, and -0.0128 $[-0.0320, 0.0028]$ for Mistral under a separate 10,000-draw source-cluster bootstrap (seed 20260925). All three intervals include zero. FEVER was not included in this earlier source-equal reweighting check; its candidate-level endpoints are reported separately below using the later endpoint-specific reconciliation. No FEVER equal-source estimate is implied. This retrospective sensitivity does not establish equivalence, semantic support validity, or a ranking advantage.

R.1 FEVER ENDPOINT SEPARATION

A follow-up audit resolved the apparent score conflict without equating `NOT_ENOUGH_INFO` with abstention. The stored `correct` field uses answer-only label accuracy: an abstaining output is not a correct answered class prediction, even when its label field matches the gold class. The separate frozen task-policy endpoint can still count an abstention as successful when `expected_action=abstain`. We therefore report class-answer yield, coverage, conditional accuracy among answers, action concordance, and the frozen composite endpoint separately.

Regime	N	Sources	Filter Δ success	Confidence tie-break Δ success
Qwen HotpotQA	468	78	$+0.0278$ $[0.0000, 0.0577]$	0.0000 $[0.0000, 0.0000]$
Qwen SciFact	202	202	$+0.0594$ $[0.0248, 0.0990]$	-0.0050 $[-0.0198, 0.0099]$
Mistral HotpotQA	222	37	$+0.0495$ $[0.0090, 0.0946]$	0.0000 $[0.0000, 0.0000]$

Table 24: Pointwise source-cluster bootstrap intervals (10,000 replicates; seed 20260919) for retrospective finite-pool task-policy success contrasts. Filter is TAFER minus the same-key unconstrained selector. Confidence tie-break is TAFER minus feasible cost/ID-only selection. Intervals are not multiplicity-adjusted or population guarantees.

Regime	TAFER task success	TAFER coverage	TAFER–uniform-feasible Δ success
Qwen HotpotQA	0.2329	0.8333	-0.0087 $[-0.0229, 0.0065]$
Qwen SciFact	0.5050	0.7327	$+0.0046$ $[-0.0149, 0.0252]$
Mistral HotpotQA	0.1892	0.7748	-0.0128 $[-0.0317, 0.0031]$

Table 25: Absolute TAFER task-policy success and answer coverage, alongside the paired TAFER-minus-uniform-feasible task-success contrast. Uniform feasible is an analytic expectation on the same feasible set. Pointwise source-cluster bootstrap intervals use 10,000 replicates and seed 20260919.

Selector	Correct-answer yield	Coverage	Answered accuracy	Action concordance	Composite task success
TAFER exact	0.4033	0.4886	0.8255	0.8628	0.8254
Same-key unconstrained	0.5655	0.7422	0.7619	0.6881	0.5988
Feasible cost/ID	0.4033	0.4886	0.8255	0.8628	0.8254
Uniform feasible	0.4068	0.4886	0.8326	0.8628	0.8278

Table 26: FEVER selector endpoints on the same frozen candidate pool (481 unique instances, 105 source clusters). Correct-answer yield uses all instances as the denominator; answered accuracy conditions on a non-abstaining selected output. Action concordance compares the abstention flag with the frozen expected action. Composite task success uses the frozen row policy and citation-ID support proxy.

Relative to the same-key unconstrained selector, TAFER increases frozen task-policy success by 0.2266 [0.1886, 0.2637] and action concordance by 0.1746 [0.1253, 0.2213], while reducing coverage by 0.2536 [-0.2964, -0.2121] and correct-answer yield by 0.1622 [-0.2037, -0.1225]. Conditional answered accuracy increases by 0.0636 [0.0064, 0.1234]. Against analytic uniform-feasible selection, the task-policy-success difference is -0.0024 [-0.0106, 0.0026], providing no evidence of an incremental selector advantage; the feasible cost/ID selector has the same values at the reported precision.

The row-level reconciliation linked 2,963 stored score rows to 1,841 pool candidate keys; 1,045 of 2,886 pool keys had no stored score row and were recomputed with the frozen evaluator and post-generation gold. All linked stored `correct` values matched answer-only EM. In 283 abstaining rows, the raw class-label field matched gold but answer-only EM was false. Stored `task_success`, `supported`, and `false_green` values matched the corresponding frozen row-policy recomputation on linked rows. Thus the earlier discrepancy was an endpoint-definition mismatch, not a reason to equate an NEI class with abstention. These estimates remain retrospective and use a citation-ID support proxy, not human support truth.

The contrast intervals are descriptive, pointwise, and retrospective. This analysis does not establish semantic support, human calibration, a causal mechanism, or generalization beyond the frozen pools. It does not replace frozen paper tables or earlier complete-case estimates.

R.2 2WIKI SAME-POOL SELECTOR RECONCILIATION

This separate 2Wiki replay links the frozen gold-free candidate pool to its post-generation exact-contract evaluation artifact. It covers 400 instances, 80 source clusters, and 2,400 candidate rows. The resolved run configuration records $K = 4$, but the selector audit contains six method-tagged records per instance; these records are not an ordered nested-$K = 4$ pool. The stored TAFER and cost-only selector outputs reproduce exactly on all 400 instances. Gold answers, expected actions, and required support IDs are joined only for evaluation, after candidate generation and selector choice.

For answer-required instances, a candidate is good when its answer is task correct and its nonempty gold-required support-ID set G is a subset of its visible cited-ID set Q. A good candidate is certified-good if it also passes the frozen operational verifier and certificate used for selection. This explicit ID contract is not natural-language entailment or human support truth. The source-cluster bootstrap is instance-weighted and pointwise (10,000 replicates; seed 20260919).

The pool has a task-correct, contract-supported candidate for 17/160 answer-required instances, but the frozen TAFER selector succeeds on 7/160. The remaining 10 available good instances are not selected successfully in this finite pool. The accounting is descriptive: it does not identify a causal effect of generation, certification, or ranking.

Across all 400 instances, feasibility filtering with the same tie-break raises task-policy success by 0.0200 [0.0000, 0.0425] versus same-key unconstrained selection, while reducing coverage by 0.0225 [-0.0425, -0.0075]. Within the same feasible set, TAFER's confidence tie-break does not improve task success over feasible cost/ID selection (-0.0025 [-0.0075, 0.0000]); it also does not exceed analytic uniform-feasible selection (-0.0038 [-0.0134, 0.0058]). These are pointwise source-cluster intervals; the audit does not support a ranking advantage.

Finite-pool term	Count / n	Estimate	95% source-cluster CI
Candidate availability A	17/160	0.1063	[0.0438, 0.1750]
Certified-good availability C	17/160	0.1063	[0.0438, 0.1750]
Candidate unavailability $1 - A$	143/160	0.8938	[0.8250, 0.9563]
Certificate exclusion $A - C$	0/160	0.0000	[0.0000, 0.0000]
Ranking regret $C - S$	10/160	0.0625	[0.0188, 0.1125]
Selected success S	7/160	0.0438	[0.0063, 0.0875]

Table 27: Answer-required 2Wiki finite-pool accounting (160 instances, 80 sources). The terms obey $1 - S = (1 - A) + (A - C) + (C - S)$ with residual 0. The point estimates exactly reconcile to the prior V40 scope receipt at absolute tolerance 10^{-12}. Intervals are descriptive pointwise source-cluster bootstrap intervals, not population guarantees.

Policy	Task success	Coverage	Correct-answer yield	Supported-answer yield	FGR_{ID}
TAFER exact	0.2275	0.7050	0.2300	0.0525	0.2125
Same-key unconstrained	0.2075	0.7275	0.2550	0.0475	0.2375
Feasible cost/ID	0.2300	0.7050	0.2300	0.0550	0.2100
Uniform feasible	0.2313	0.7050	0.2386	0.0567	0.2174

Table 28: All-instance policy outcomes on the same 400 instances and 80 sources. FGR_{ID} is a gold citation-ID proxy under the explicit contract, not a human-calibrated semantic false-green rate. Uniform feasible is the analytic mean over the same feasible set.

The direction of the exact-contract set relation matters. Among 679 answer-required, non-abstaining candidate rows, the frozen support relation $G \neq \varnothing, G \subseteq Q$ and the distinct precision-style relation $Q \neq \varnothing, Q \subseteq G$ agree on 116 rows; 7 satisfy only the former, 279 only the latter, and 277 neither. The 286 disagreements are differences between ID-set constructs, not estimates of independent checker error or semantic-support validity.

S SOURCE-WEIGHTING SENSITIVITY FOR THE FRESH V87 DIAGNOSTIC

The primary V87 contrasts weight unique instances equally and resample their shared source clusters. As a weighting sensitivity, we first average paired outcomes within each observed source and then give each source equal weight. This changes the estimand; it is not a replacement for the frozen instance-weighted analysis. Intervals are pointwise source-cluster percentile bootstrap intervals with 2,000 draws and seed 20260919. Every endpoint retains its own complete-case denominator, shown as observed/planned/source clusters.

For objective task behavior, source-equal weighting leaves the direction of the pre-specified uncited-removal minus cited-removal TextEM contrast unchanged in all three domains. The 2Wiki interval still includes zero. FEVER's point estimate is larger under equal-source weighting, so its magnitude depends on the weighting estimand even though both estimates are positive. Coverage and AnsweredEM are reported separately; FEVER's task label `NOT ENOUGH INFO` is not treated as abstention (Table 29).

Operational-judge quantities remain proxies, not human support truth (Table 30). For the primary contrast, their direction is stable across the two weighting schemes. Across the full secondary contrast grid, one small HotpotQA unsupported-assertion proxy contrast against the sham condition changes sign under equal-source weighting; we retain that discrepancy as a sensitivity finding, not as evidence for a directional mechanism. The full 135-row endpoint/contrast grid, missingness, input hashes, and exact estimates are in the accompanying machine-readable receipt. Two no-evidence support-proxy contrasts have no observed complete pairs and remain unestimated.

Domain	Endpoint	$n/N/S$	Instance-weighted	Equal-source
HotpotQA	TextEM	100/100/62	+0.170 [+0.082, +0.270]	+0.161 [+0.073, +0.258]
HotpotQA	AnsweredEM	100/100/62	+0.170 [+0.082, +0.270]	+0.161 [+0.073, +0.266]
HotpotQA	Coverage	100/100/62	+0.330 [+0.224, +0.441]	+0.290 [+0.194, +0.387]
FEVER	TextEM	80/80/67	+0.575 [+0.430, +0.718]	+0.672 [+0.552, +0.791]
FEVER	AnsweredEM	80/80/67	+0.575 [+0.430, +0.718]	+0.672 [+0.552, +0.791]
FEVER	Coverage	80/80/67	+0.000 [+0.000, +0.000]	+0.000 [+0.000, +0.000]
2Wiki	TextEM	100/100/60	+0.100 [-0.010, +0.222]	+0.092 [-0.008, +0.200]
2Wiki	AnsweredEM	100/100/60	+0.100 [-0.010, +0.222]	+0.092 [+0.000, +0.192]
2Wiki	Coverage	100/100/60	+0.130 [+0.020, +0.238]	+0.142 [+0.042, +0.250]

Table 29: Instance-weighted and equal-source paired differences for objective task behavior under uncited removal minus cited removal. Each row uses its own complete-case denominator. $n/N/S$ denotes observed pairs, planned pairs, and observed source clusters. Intervals are pointwise 95% source-cluster percentile bootstrap intervals (2,000 draws; seed 20260919).

Domain	Endpoint	$n/N/S$	Instance-weighted	Equal-source
HotpotQA	Support proxy	57/100/43	+0.228 [+0.088, +0.382]	+0.221 [+0.081, +0.360]
HotpotQA	Supported answer proxy	100/100/62	+0.370 [+0.252, +0.486]	+0.363 [+0.242, +0.484]
HotpotQA	Unsupported assertion proxy	100/100/62	-0.040 [-0.162, +0.085]	-0.073 [-0.210, +0.056]
HotpotQA	Checker-FGR proxy	100/100/62	-0.040 [-0.100, +0.010]	-0.048 [-0.113, +0.000]
FEVER	Support proxy	80/80/67	+0.650 [+0.524, +0.773]	+0.709 [+0.590, +0.821]
FEVER	Supported answer proxy	80/80/67	+0.650 [+0.524, +0.773]	+0.709 [+0.590, +0.821]
FEVER	Unsupported assertion proxy	80/80/67	-0.650 [-0.773, -0.524]	-0.709 [-0.821, -0.582]
FEVER	Checker-FGR proxy	80/80/67	-0.013 [-0.096, +0.063]	+0.000 [-0.090, +0.075]
2Wiki	Support proxy	48/100/32	+0.208 [+0.060, +0.362]	+0.188 [+0.062, +0.328]
2Wiki	Supported answer proxy	100/100/60	+0.200 [+0.096, +0.309]	+0.192 [+0.092, +0.292]
2Wiki	Unsupported assertion proxy	100/100/60	-0.070 [-0.176, +0.039]	-0.050 [-0.158, +0.050]
2Wiki	Checker-FGR proxy	100/100/60	-0.010 [-0.120, +0.109]	-0.008 [-0.108, +0.092]

Table 30: Instance-weighted and equal-source paired differences for operational-judge endpoints under uncited removal minus cited removal. These are proxies, not human support truth. $n/N/S$ denotes observed pairs, planned pairs, and observed source clusters. Intervals are pointwise 95% source-cluster percentile bootstrap intervals (2,000 draws; seed 20260919).

T SINGLE-SOURCE INFLUENCE FOR THE FRESH V87 CONTRAST

We recomputed the frozen uncited-removal minus cited-removal contrast after deleting each observed source cluster in turn. Table 31 shows that the objective TextEM point estimate stays positive in each domain under every single-source deletion. This influence check is descriptive: its ranges are not confidence intervals and do not establish population-level robustness. The estimates use each endpoint's own complete cases.

The operational checker-FGR contrast is less stable (Table 32). Its leave-one-source range crosses zero for FEVER and 2Wiki; the FEVER equal-source full estimate is exactly zero, so its direction is undefined. These are operational-proxy diagnostics, not human support-validity results. The source-deletion analysis neither changes the frozen cohorts nor closes the human-validity route.

The finite set of observed source clusters is small enough that deleting one cluster can materially change some operational-proxy contrasts. We therefore retain the proxy sign sensitivity rather than describing the checker result as source-invariant. No source IDs or additional samples are introduced by this analysis.

We also checked whether the response-mode crosswalk depends on one source. On the frozen common four-condition cohort, we compared the sham-repeat condition with cited-evidence removal, weighting source-level paired means equally and deleting every source once. Positive TextEM/AnsweredEM values mean a larger task-correctness loss under cited removal; positive abstention and conflict values mean higher rates under cited removal. Table 33 reports the source-equal point estimate and the range across all single-source deletions. These ranges are descriptive influence summaries, not confidence intervals. HotpotQA's point direction is unchanged under every deletion; 2Wiki's source-equal point direction is also unchanged, while its previously reported instance-weighted source-cluster interval still includes zero. FEVER TextEM and AnsweredEM remain unavailable in this scored-row artifact; its abstention contrasts are zero in this cohort. This sensitivity does not establish natural-language support validity.

Table 31: Leave-one-source influence for objective task TextEM on the frozen V87 primary contrast. Each range is the minimum and maximum point estimate after deleting one observed source cluster; it is not an interval estimate.

Domain	n/sources	Instance full	Instance LOO range	Equal-source full	Equal-source LOO range
HotpotQA	100/62	+0.170	[+0.153,+0.182]	+0.161	[+0.148,+0.180]
FEVER	80/67	+0.575	[+0.564,+0.615]	+0.672	[+0.667,+0.697]
2Wiki	100/60	+0.100	[+0.082,+0.122]	+0.092	[+0.076,+0.110]

Table 32: Leave-one-source influence for the operational checker-FGR contrast. Values are operational proxies; ranges are descriptive single-source-deletion point estimates, not confidence intervals.

Domain	n/sources	Instance full	Instance LOO range	Equal-source full	Equal-source LOO range
HotpotQA	100/62	-0.040	[-0.051,-0.020]	-0.048	[-0.057,-0.033]
FEVER	80/67	-0.013	[-0.025,+0.013]	+0.000	[-0.015,+0.015]
2Wiki	100/60	-0.010	[-0.031,+0.010]	-0.008	[-0.025,+0.008]

Table 33: Source-equal response-mode estimates and leave-one-source-out (LOO) ranges for cited removal versus sham repeat. Ranges are descriptive, not confidence intervals. FEVER's generic TextEM and AnsweredEM are unavailable.

Domain	n/src.	TextEM / AnsweredEM	Abstain	Flag conflict
HotpotQA	99/62	$+.169$ [$+.156, +.189$]	$+.298$ [$+.287, +.303$]	$+.024$ [$+.016, +.025$]
FEVER	80/67	NA	$+.000$ [$+.000, +.000$]	$+.000$ [$+.000, +.000$]
2Wiki	100/60	$+.092$ [$+.076, +.110$]	$+.133$ [$+.119, +.153$]	$+.008$ [$+.000, +.008$]